

Broadview
博文视点

涵盖并行程序设计基础、思路、方法和实战
内容丰富，实例典型，实用性强

Broadview®
www.broadview.com.cn



Java 高并发程序设计

葛一鸣 郭超 编著

中国工信出版集团

电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

实战
Java高并发程序设计

中国工业出版社
000000

Java 高并发程序设计

葛一鸣 郭超 编著

电子工业出版社
Publishing House of Electronics Industry
北京•BEIJING

目 录

[内容简介](#)

[前言](#)

[第1章 走入并行世界](#)

[1.1 何去何从的并行计算](#)

[1.1.1 忘掉那该死的并行](#)

[1.1.2 可怕的现实：摩尔定律的失效](#)

[1.1.3 柳暗花明：不断地前进](#)

[1.1.4 光明或是黑暗](#)

[1.2 你必须知道的几个概念](#)

[1.2.1 同步（Synchronous）和异步（Asynchronous）](#)

[1.2.2 并发（Concurrency）和并行（Parallelism）](#)

[1.2.3 临界区](#)

[1.2.4 阻塞（Blocking）和非阻塞（Non-Blocking）](#)

[1.2.5 死锁（Deadlock）、饥饿（Starvation）和活锁（Livelock）](#)

[1.3 并发级别](#)

[1.3.1 阻塞（Blocking）](#)

[1.3.2 无饥饿（Starvation-Free）](#)

[1.3.3 无障碍（Obstruction-Free）](#)

[1.3.4 无锁（Lock-Free）](#)

[1.3.5 无等待（Wait-Free）](#)

[1.4 有关并行的两个重要定律](#)

[1.4.1 Amdahl定律](#)

[1.4.2 Gustafson定律](#)

[1.4.3 Amdahl定律和Gustafson定律是否相互矛盾](#)

[1.5 回到Java: JMM](#)

[1.5.1 原子性 \(Atomicity\)](#)

[1.5.2 可见性 \(Visibility\)](#)

[1.5.3 有序性 \(Ordering\)](#)

[1.5.4 哪些指令不能重排: Happen-Before规则](#)

[1.6 参考文献](#)

[第2章 Java并程序基础](#)

[2.1 有关线程你必须知道的事](#)

[2.2 初始线程: 线程的基本操作](#)

[2.2.1 新建线程](#)

[2.2.2 终止线程](#)

[2.2.3 线程中断](#)

[2.2.4 等待 \(wait\) 和通知 \(notify\)](#)

[2.2.5 挂起 \(suspend\) 和继续执行 \(resume\) 线程](#)

[2.2.6 等待线程结束 \(join\) 和谦让 \(yield\)](#)

[2.3 volatile与Java内存模型 \(JMM\)](#)

[2.4 分门别类的管理: 线程组](#)

[2.5 驻守后台: 守护线程 \(Daemon\)](#)

[2.6 先干重要的事: 线程优先级](#)

[2.7 线程安全的概念与synchronized](#)

[2.8 程序中的幽灵: 隐蔽的错误](#)

[2.8.1 无提示的错误案例](#)

[2.8.2 并发下的ArrayList](#)

[2.8.3 并发下诡异的HashMap](#)

[2.8.4 初学者常见问题: 错误的加锁](#)

[2.9 参考文献](#)

[第3章 JDK并发包](#)

[3.1 多线程的团队协作：同步控制](#)

[3.1.1 synchronized的功能扩展：重入锁](#)

[3.1.2 重入锁的好搭档：Condition条件](#)

[3.1.3 允许多个线程同时访问：信号量（Semaphore）](#)

[3.1.4 ReadWriteLock读写锁](#)

[3.1.5 倒计时器：CountDownLatch](#)

[3.1.6 循环栅栏：CyclicBarrier](#)

[3.1.7 线程阻塞工具类：LockSupport](#)

[3.2 线程复用：线程池](#)

[3.2.1 什么是线程池](#)

[3.2.2 不要重复发明轮子：JDK对线程池的支持](#)

[3.2.3 刨根究底：核心线程池的内部实现](#)

[3.2.4 超负载了怎么办：拒绝策略](#)

[3.2.5 自定义线程创建：ThreadFactory](#)

[3.2.6 我的应用我做主：扩展线程池](#)

[3.2.7 合理的选择：优化线程池线程数量](#)

[3.2.8 堆栈去哪里了：在线程池中寻找堆栈](#)

[3.2.9 分而治之：Fork/Join框架](#)

[3.3 不要重复发明轮子：JDK的并发容器](#)

[3.3.1 超好用的工具类：并发集合简介](#)

[3.3.2 线程安全的HashMap](#)

[3.3.3 有关List的线程安全](#)

[3.3.4 高效读写的队列：深度剖析ConcurrentLinkedQueue](#)

[3.3.5 高效读取：不变模式下的CopyOnWriteArrayList](#)

[3.3.6 数据共享通道：BlockingQueue](#)

[3.3.7 随机数据结构：跳表（SkipList）](#)

[3.4 参考资料](#)

[第4章 锁的优化及注意事项](#)

[4.1 有助于提高“锁”性能的几点建议](#)

[4.1.1 减小锁持有时间](#)

[4.1.2 减小锁粒度](#)

[4.1.3 读写分离锁来替换独占锁](#)

[4.1.4 锁分离](#)

[4.1.5 锁粗化](#)

[4.2 Java虚拟机对锁优化所做的努力](#)

[4.2.1 锁偏向](#)

[4.2.2 轻量级锁](#)

[4.2.3 自旋锁](#)

[4.2.4 锁消除](#)

[4.3 人手一支笔：ThreadLocal](#)

[4.3.1 ThreadLocal的简单使用](#)

[4.3.2 ThreadLocal的实现原理](#)

[4.3.3 对性能有何帮助](#)

[4.4 无锁](#)

[4.4.1 与众不同的并发策略：比较交换（CAS）](#)

[4.4.2 无锁的线程安全整数：AtomicInteger](#)

[4.4.3 Java中的指针：Unsafe类](#)

[4.4.4 无锁的对象引用：AtomicReference](#)

[4.4.5 带时间戳的对象引用：AtomicStampedReference](#)

[4.4.6 数组也能无锁：AtomicIntegerArray](#)

[4.4.7 让普通变量也享受原子操作：AtomicIntegerFieldUpdater](#)

[4.4.8 挑战无锁算法：无锁的Vector实现](#)

[4.4.9 让线程之间互相帮助：细看SynchronousQueue的实现](#)

[4.5 有关死锁的问题](#)

[4.6 参考文献](#)

[第5章 并行模式与算法](#)

[5.1 探讨单例模式](#)

[5.2 不变模式](#)

[5.3 生产者-消费者模式](#)

[5.4 高性能的生产者-消费者：无锁的实现](#)

[5.4.1 无锁的缓存框架：Disruptor](#)

[5.4.2 用Disruptor实现生产者-消费者案例](#)

[5.4.3 提高消费者的响应时间：选择合适的策略](#)

[5.4.4 CPU Cache的优化：解决伪共享问题](#)

[5.5 Future模式](#)

[5.5.1 Future模式的主要角色](#)

[5.5.2 Future模式的简单实现](#)

[5.5.3 JDK中的Future模式](#)

[5.6 并行流水线](#)

[5.7 并行搜索](#)

[5.8 并行排序](#)

[5.8.1 分离数据相关性：奇偶交换排序](#)

[5.8.2 改进的插入排序：希尔排序](#)

[5.9 并行算法：矩阵乘法](#)

[5.10 准备好了再通知我：网络NIO](#)

[5.10.1 基于Socket的服务端的多线程模式](#)

[5.10.2 使用NIO进行网络编程](#)

[5.10.3 使用NIO来实现客户端](#)

[5.11 读完了再通知我：AIO](#)

[5.11.1 AIO EchoServer的实现](#)

[5.11.2 AIO Echo客户端实现](#)

[5.12 参考文献](#)

[第6章 Java 8与并发](#)

[6.1 Java 8的函数式编程简介](#)

[6.1.1 函数作为一等公民](#)

[6.1.2 无副作用](#)

[6.1.3 申明式的（Declarative）](#)

[6.1.4 不变的对象](#)

[6.1.5 易于并行](#)

[6.1.6 更少的代码](#)

[6.2 函数式编程基础](#)

[6.2.1 FunctionalInterface注释](#)

[6.2.2 接口默认方法](#)

[6.2.3 lambda表达式](#)

[6.2.4 方法引用](#)

[6.3 一步一步走入函数式编程](#)

[6.4 并行流与并行排序](#)

[6.4.1 使用并行流过滤数据](#)

[6.4.2 从集合得到并行流](#)

[6.4.3 并行排序](#)

[6.5 增强的Future：CompletableFuture](#)

[6.5.1 完成了就通知我](#)

[6.5.2 异步执行任务](#)

[6.5.3 流式调用](#)

[6.5.4 CompletableFuture中的异常处理](#)

[6.5.5 组合多个CompletableFuture](#)

[6.6 读写锁的改进：StampedLock](#)

[6.6.1 StampedLock使用示例](#)

[6.6.2 StampedLock的小陷阱](#)

[6.6.3 有关StampedLock的实现思想](#)

[6.7 原子类的增强](#)

[6.7.1 更快的原子类：LongAdder](#)

[6.7.2 LongAdder的功能增强版：LongAccumulator](#)

[6.8 参考文献](#)

[第7章 使用Akka构建高并发程序](#)

[7.1 新并发模型：Actor](#)

[7.2 Akka之Hello World](#)

[7.3 有关消息投递的一些说明](#)

[7.4 Actor的生命周期](#)

[7.5 监督策略](#)

[7.6 选择Actor](#)

[7.7 消息收件箱（Inbox）](#)

[7.8 消息路由](#)

[7.9 Actor的内置状态转换](#)

[7.10 询问模式：Actor中的Future](#)

[7.11 多个Actor同时修改数据：Agent](#)

[7.12 像数据库一样操作内存数据：软件事务内存](#)

[7.13 一个有趣的例子：并发粒子群的实现](#)

[7.13.1 什么是粒子群算法](#)

[7.13.2 粒子群算法的计算过程](#)

[7.13.3 粒子群算法能做什么](#)

[7.13.4 使用Akka实现粒子群](#)

[7.14 参考文献](#)

[第8章 并程序调试](#)

[8.1 准备实验样本](#)

[8.2 正式起航](#)

[8.3 挂起整个虚拟机](#)

[8.4 调试进入ArrayList内部](#)

图书在版编目（**CIP**）数据

实战Java高并发程序设计 / 葛一鸣，郭超编著. ——北京：电子工业出版社，2015.11

ISBN 978-7-121-27304-9

I. ①实... II. ①葛... ②郭... III. ①JAVA语言—程序设计 IV. ①TP312

中国版本图书馆CIP数据核字（2015）第231507号

责任编辑：董 英

印 刷：

装 订：

出版发行：电子工业出版社

北京市海淀区万寿路173信箱

邮 编：100036

开 本：787×980 1/16

印 张：22

字 数：493千字

版 次：2015年11月第1版

印 次：2015年11月第1次印刷

印 数：3000册

定 价：69.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线：（010）88258888。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

内容简介

在过去单核CPU时代，单任务在一个时间点只能执行单一程序，随着多核CPU的发展，并行程序开发就显得尤为重要。

本书主要介绍基于Java的并行程序设计基础、思路、方法和实战。第一，立足于并发程序基础，详细介绍Java中进行并行程序设计的基本方法。第二，进一步详细介绍JDK中对并行程序的强大支持，帮助读者快速、稳健地进行并行程序开发。第三，详细讨论有关“锁”的优化和提高并行程序性能级别的方法和思路。第四，介绍并行的基本设计模式及Java 8对并行程序的支持和改进。第五，介绍高并发框架Akka的使用方法。最后，详细介绍并行程序的调试方法。

本书内容丰富，实例典型，实用性强，适合有一定Java基础的技术开发人员阅读。

前言

关于Java与并行

由于单核CPU的主频逐步逼近极限，多核CPU架构成为了一种必然的技术趋势。所以，多线程并行程序便显得越来越重要。并行计算的一个重要应用场景就是服务端编程。可以看到，目前服务端CPU的核心数已经轻松超越10核心，而Java显然已经成为当下最流行的服务端编程语言，因此熟悉和了解基于Java的并行程序开发有着重要的实用价值。

本书的体系结构

本书立足于实际开发，又不缺乏理论介绍，力求通俗易懂、循序渐进。本书共分为8章。

第1章主要介绍了并行计算中相关的一些基本概念，树立读者对并行计算的基本认识；介绍了两个重要的并行性能评估定律，以及Java内存模型JMM。

第2章介绍了Java并行程序开发的基础，包括Java中Thread的基本使用方法等，也详细介绍了并行程序容易引发的一些错误和误用。

第3章介绍了JDK内部对并行程序开发的支持，主要介绍JUC（Java.util.concurrent）中一些工具的使用方法、各自特点及它们的

内部实现原理。

第4章介绍了在开发过程中可以进行的对锁的优化，也进一步简要描述了Java虚拟机层面对并程序序的优化支持。此外，还花费一定篇幅介绍了有关无锁的计算。

第5章介绍了并程序序设计中常见的一些设计模式以及一些典型的并行算法和使用方法，其中包括重要的Java NIO和AIO的介绍。

第6章介绍了Java 8中为并行计算做的新的改进，包括并行流、CompletableFuture、StampedLock和LongAdder。

第7章主要介绍了高并发框架Akka的基本使用方法，并使用Akka框架实现了一个简单的粒子群算法，模拟超高并发的场景。

第8章介绍了使用Eclipse进行多线程调试的方法，并演示了通过Eclipse进行多线程调试重现ArrayList的线程不安全问题。

本书特色

本书的主要特点如下。

1. 结构清晰。本书一共8章，总体上循序渐进，逐步提升。每一章都各自有鲜明的侧重点，有利于读者快速抓住重点。
2. 理论结合实战。本书注重实战，书中重要的知识点都安排了代码实例，帮助读者理解。同时也不忘记对系统的内部实现原理进行深度剖析。
3. 通俗易懂。本书尽量避免采用过于理论的描述方式，简单的白话

文风格贯穿全书，配图基本上为手工绘制，降低了理解难度，并尽量做到读者在阅读过程中少盲点、无盲点。

适合阅读人群

虽然本书力求通俗，但要通读本书并取得良好的学习效果，要求读者需要具备基本的Java知识或者一定的编程经验。因此，本书适合以下读者：

- 拥有一定开发经验的Java平台开发人员（Java、Scala、JRuby等）
- 软件设计师、架构师
- 系统调优人员
- 有一定的Java编程基础并希望进一步加深对并行的理解的研发人员

本书的约定

本书在叙述过程中，有如下约定：

- 本书中所述的JDK 1.5、JDK 1.6、JDK 1.7、JDK 1.8分别等同于JDK 5、JDK 6、JDK 7、JDK 8。
- 如无特殊说明，本书的程序、示例均在JDK 1.7环境中运行。

联系作者

本书的写作过程远比我想象得更艰辛，为了让全书能够更清楚、更正确地表达和论述，我经历了很多个不眠之夜，即使现在回想起来，我也忍不住会打个寒战。由于写作水平的限制，书中难免会有不妥之处，望读者谅解。

为此，如果读者有任何疑问或者建议，非常欢迎大家加入QQ群397196583，一起探讨学习中的困难、分享学习的经验，我期待与大家一起交流、共同进步。同时，也希望大家可以关注我的博客<http://www.uucode.net/>。

感谢

这本书能够面世，是因为得到了众人的支持。首先，要感谢我的妻子，她始终不辞辛劳、毫无怨言地对我照顾有加，才让我得以腾出大量时间，并可以安心工作。其次，要感谢所有编辑为我一次又一次地审稿改错，批评指正，才能让本书逐步完善。最后，感谢我的母亲30年如一日对我的体贴和关心。

参与本书编写的还有安继宏、白慧、薛淑英、蒋玺、曹静、马玉杰、陈明明、张丽萍、任娜娜、李清艺、荆海霞、赵全利、孙迪，在此一并感谢！

葛一鸣

第1章 走入并行世界

当你打开本书，也许你正试图将你的应用改造成并行模式运行，也许你只是单纯地对并行程序感兴趣。无论出于何种原因，你正对并行计算充满好奇、疑问和求知欲。如果是这样，那就对了，带着你的好奇和疑问，让我们一起遨游并行程序的世界，深入了解它们究竟是如何工作的吧！

不过首先，我想要公布一条令人沮丧的消息。就在大伙儿都认为并行计算必然成为未来的大趋势时，2014年底，Avoiding ping pong论坛上，伟大的Linus Torvalds提出了一个截然不同的观点，他说：“忘掉那该死的并行吧！”（原文：Give it up. The whole "parallel computing is the future" is a bunch of crock.）

1.1 何去何从的并行计算

到底我们该如何选择呢？本节的目的就是拨云见日。

1.1.1 忘掉那该死的并行

Linus Torvalds是一个传奇式的人物（图1.1），是他给出了Linux的原型，并一直致力于推广和发展Linux系统。他在1991年首先在网络上发布了Linux源码，从此一发而不可收。Linux迅速崛起壮大，成为目前使用最广泛的操作系统之一。



图1-1 传奇的Linus Torvalds

自2002年起，Linus就决定使用BitKeeper作为Linux内核开发的版本控制工具，以此来维护Linux的内核源码。BitKeeper是一套分布式版本控制软件，它是一套商用系统，由BitMover公司开发。2005年，BitKeeper宣称发现Linux内核开发人员使用逆向工程来试图解析BitKeeper内部协议。因此，决定向Linus收回BitKeeper授权。尽管Linux核心团队与BitMover公司进行了协商，但是无法解决他们之间的分歧。因此，Linus决定自行研发版本控制工具来代替BitKeeper。于是，Git诞生了。

如果大家正在使用Git，我相信你们一定会被Git的魅力所折服，如果还没有了解过Git，那么我强烈建议你去关注一下这款优秀的产品。

而正是这位传奇人物，给目前红红火火的并行计算泼了一大盆冷水。那么，并行计算究竟应该何去何从呢？

在Linus的发言中这么说道：

Where the hell do you envision that those magical parallel algorithms would be used?

The only place where parallelism matters is in graphics or on the server side, where we already largely have it. Pushing it anywhere else is just pointless.

需要有多么奇葩的想象力才能想象出并行计算的用武之地？

并行计算只有在图像处理和服务端编程2个领域可以使用，并且它在这2个领域确实有着大量广泛的使用。但是在其他任何地方，并行计算毫无建树！

So the whole argument that people should parallelize their code is fundamentally flawed. It rests on incorrect assumptions. It's a fad that has been going on too long.

因此，人们在争论是否应该将他们的代码并行化是一个本质上的错误。这完全就基于一个错误的假设。“并行”是一个早该结束的时髦用语。

看了这段较为完整的表述，大家应该对Linus的观点有所感触，我对此也表示赞同。与串程序不同，并程序的设计和实现异常复杂，不仅仅体现在程序的功能分离上，多线程间的协调性、乱序性都会成为程序正确执行的障碍。只要你稍不留神，就会失之毫厘，谬以千里！混乱的程序难以阅读、难以理解，更难以调试。所谓并行，也就是把简单问题复杂化的典型。因此，只有“疯子”才会叫嚣并行就是未来（the crazies talking about scaling to hundreds of cores are just that - crazy）。

但是，Linus也提出了两个特例，那就是图像处理和服务端程序是可以、也需要使用并行技术的。仔细想想，为什么图像处理和服务端程序是特例呢？

和用户终端程序不同，图像处理往往拥有极大的计算量。一张1024×768像素的图片，包含多达78万6千多个像素。即使将所有的像素遍历一遍，也得花不少时间。更何况，图像处理涉及大量的矩阵计算。矩阵的规模和数量都会非常大。面对如此密集的计算，很有可能超过单核CPU的计算能力，所以自然需要引入多核计算了。

而服务端程序与一般的用户终端程序相比，一方面，服务端程序需要承受很重的用户访问压力。根据淘宝的数据，它在“双十一”一天，支

付宝核心数据库集群处理了41亿个事务，执行285亿次SQL，生成15TB日志，访问1931亿次内存数据块，13亿个物理读。如此密集访问，恐怕任何一台单机都难以胜任，因此，并行程序也就自然成了唯一的出路。另一方面，服务端程序往往会比用户终端程序拥有更复杂的业务模型。面对复杂业务模型，并行程序会比串行程序更容易适应业务需求，更容易模拟我们的现实世界。毕竟，我们的世界本质上是并行的。比如，当你开开心心去上学的时候，妈妈可能在家里忙着家务，爸爸在外打工赚钱，一家人其乐融融。如果有一天，你需要使用你的计算机来模拟这个场景，你会怎么做呢？如果你就在一个线程里，既做了你自己，又做了妈妈，又做了爸爸，显然这不是一种好的解决方案。但如果你使用三个线程，分别模拟这三个人，一切看起来又是那么自然，而且容易被人理解。

再举一个专业点的例子，比如基础平台Java虚拟机，虚拟机除了要执行main函数主线程外，还需要做JIT编译，需要做垃圾回收。无论是main函数、JIT编译还是垃圾回收，在虚拟机内部都实现为单独的一个线程。是什么使得虚拟机的研发人员这么做呢？显然，这是因为建模的需要。因为这里的每一个任务都是相对独立的。我们不应该将没有关联的业务代码拼凑在一起，分离为不同的线程更容易理解和维护。因此，使用并行也不完全出自性能的考虑，而有时候，我们会很自然地那么做。

1.1.2 可怕的现实：摩尔定律的失效

摩尔定律是由英特尔创始人之一戈登·摩尔提出来的。其内容为：集成电路上可容纳的电晶体（晶体管）数目，约每隔24个月便会增加一

倍；经常被引用的“18个月”，是由英特尔首席执行官大卫·豪斯所说：预计18个月会将芯片的性能提高一倍（即更多的晶体管使其更快）。

说得直白点，就是每18个月到24个月，我们的计算机性能就能翻一番。

反过来说，就是每过18个月到24个月，你在未来用一半的价钱就能买到和现在性能相同的计算设备了。这听起来是一件多么激动人心的事情呀！

但是，摩尔定律并不是一种自然法则或者物理定律，它只是基于人为观测数据后，对未来的预测。按照这种速度，我们的计算能力将会按照指数速度增长，用不了多久，我们的计算能力就能超越“上帝”了！畅想未来，基于强劲的超级计算机，我们甚至可以模拟整个宇宙。

摩尔定律的有效性已经超过半个世纪了，然而，在2004年，Intel宣布将4GHz芯片的发布时间推迟到2005年，在2004年秋季，Intel宣布彻底取消4GHz计划（图1.2）。



图1.2 Intel CEO Barret单膝下跪对取消4GHz感到抱歉

是什么迫使世界顶级的科技巨头放弃4GHz的研发呢？显然，就目前的硅电路而言，很有可能已经走到了头。我们的制造工艺已经到了纳米了。1纳米是 10^{-9} 米，也就是10亿分之一米。这已经是一个相当小的数字了。就目前的科技水平而言，如果无法在物质分子层面以下进行工作，那么也许4GHz的芯片就已经接近了理论极限。因为即使一个水分子，它的直径也有0.4纳米。再往下发展就显得有些困难。当然，如果我们使用完全不同的计算理论或者芯片生产工艺，也许会有本质的突破，但目前还没有看到这种技术被大规模使用的可能。

因此，摩尔定律在CPU的计算性能上可能已经失效。虽然，现在Intel已经研制出了4GHz芯片，但可以看到，在近10年的发展中，CPU主频的提升已经明显遇到了一些暂时不可逾越的瓶颈。

1.1.3 柳暗花明：不断地前进

虽然CPU的性能已经几近止步，长达半个世纪的摩尔定律轰然倒地。但是这依然没有阻挡科学家和工程师们带领我们不断向前的脚步。

从2005年开始，我们已经不再追求单核的计算速度，而着迷于研究如何将多个独立的计算单元整合到单独的CPU中，也就是我们所说的多核CPU。短短十几年的发展，家用型CPU，比如Intel i7就可以拥有4核心，甚至8核心。而专业服务器则通常可以配有几个独立的CPU，每一个CPU都拥有多达8个甚至更多的内核。从整体上看，专业服务器的内核总数甚至可以达到几百个。

非常令人激动，摩尔定律在另外一个侧面又生效了。根据这个定律，我们可以预测，每过18到24个月，CPU的核心数就会翻一番。用不了多久，拥有几十甚至上百CPU内核的芯片就能进入千家万户。

顶级计算机科学家唐纳德·尔文·克努斯（Donald Ervin Knuth），如此评价这种情况：在我看来，这种现象（并发）或多或少是由于硬件设计者已经无计可施了导致的，他们将摩尔定律失效的责任推脱给软件开发人员。

唐纳德（图1.3）是著名计算机巨著《计算机程序设计艺术》的作者。《美国科学家》杂志曾将该书与爱因斯坦的《相对论》，狄拉克的《量子力学》和理查·费曼的《量子电动力学》等书并列为20世纪最重要的12本物理科学类专论书之一。

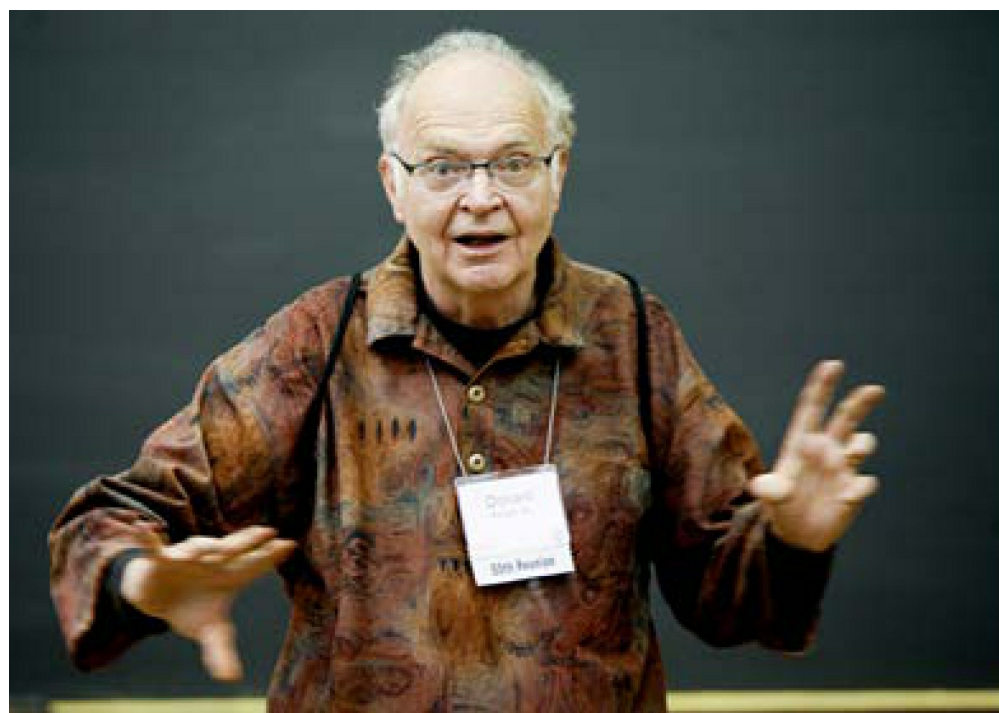


图1.3 唐纳德院士

1.1.4 光明或是黑暗

根据唐纳德的观点，摩尔定律本应该由硬件开发人员维持。但是，很不幸，硬件工程师似乎已经无计可施了。为了继续保持性能的高速发展，硬件工程师就破天荒地想出了将多个CPU内核塞进一个CPU里的奇妙想法。由此，并行计算就被非常自然地推广开来，而随之而来的问题也层出不穷，程序员的黑暗时期也随之到来。简化的硬件设计方案必然带来软件设计的复杂性。换句话说，软件工程师正在为硬件工程师无法完成的工作负责，因此，也就有了唐纳德的“他们将摩尔定律失效的责任推脱给了软件开发者的说法。

所以，如何让多个CPU有效并且正确地工作也就成为了一门技术，甚至是很大的学问。比如，多线程间如何保证线程安全，如何正确理解线程间的无序性、可见性，如何尽可能提高并行程序的设计，又如何将串行程序改造为并行程序。而对并行计算的研究，也就是希望在这片黑暗中带来光明。

1.2 你必须知道的几个概念

现在，并行计算显然已经成为一门正式的学问。也许很多人（包括Linux在内），都会觉得并行计算或者说并行算法是多么奇葩。但现在我们也不得不承认，在某些领域，这些算法还是有用武之地的。既然说服务端编程还是大量需要并行计算的，而Java也主要占领着服务端市场，那么对Java的并行计算的研究也就显得非常的必要。但首先，我想在这里先介绍几个重要的相关概念。

1.2.1 同步（**Synchronous**）和异步（**Asynchronous**）

同步和异步通常用来形容一次方法调用。同步方法调用一旦开始，调用者必须等到方法调用返回后，才能继续后续的行为。异步方法调用更像一个消息传递，一旦开始，方法调用就会立即返回，调用者就可以继续后续的操作。而异步方法通常会在另外一个线程中“真实”地执行。整个过程，不会阻碍调用者的工作。图1.4显示了同步方法调用和异步方法调用的区别。对于调用者来说，异步调用似乎是一瞬间就完成的。如果异步调用需要返回结果，那么当这个异步调用真实完成时，则会通知调用者。

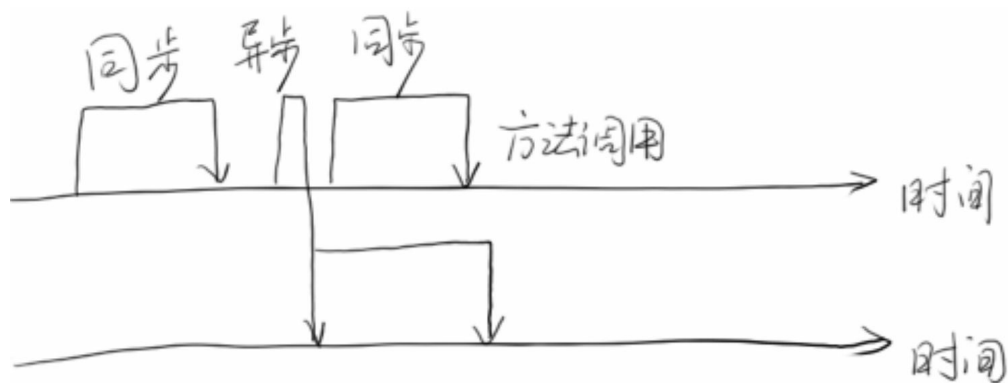


图1.4 同步和异步方法调用

打个比方，比如我们去购物，如果你去商场实体店买一台空调，当你到了商场看中了一款空调，你就想售货员下单。售货员去仓库帮你调配物品。这天你热得实在不行了，就催着商家赶紧给你送货，于是你就等在商店里，候着他们，直到商家把你和空调一起送回家，一次愉快的购物就结束了。这就是同步调用。

不过，如果我们赶时髦，就坐在家打开电脑，在网上订购了一台空调。当你完成网上支付的时候，对你来说购物过程已经结束了。虽然空调还没送到家，但是你的任务都已经完成了。商家接到了你的订单后，就会加紧安排送货，当然这一切已经跟你无关了。你已经支付完成，想干什么就能去干什么，出去溜几圈都不成问题，等送货上门的时候，接到商家的电话，回家一趟签收就完事了。这就是异步调用。

1.2.2 并发（**Concurrency**）和并行（**Parallelism**）

并发和并行是两个非常容易被混淆的概念。它们都可以表示两个或者多个任务一起执行，但是偏重点有些不同。并发偏重于多个任务交替

执行，而多个任务之间有可能还是串行的。而并行是真正意义上的“同时执行”。图1.5很好地诠释了这点。

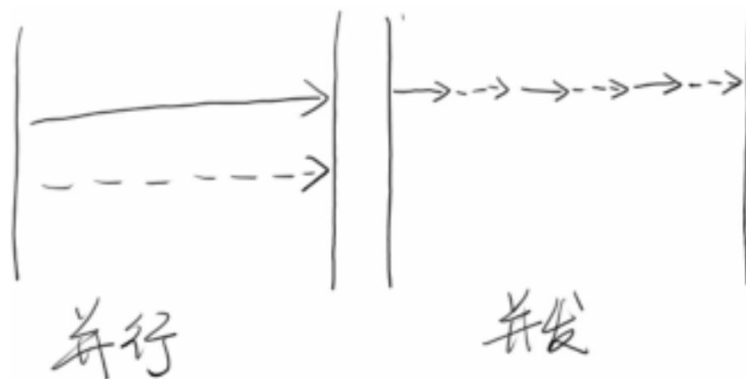


图1.5 并发和并行

严格意义上来说，并行的多个任务是真实的同时执行，而对于并发来说，这个过程只是交替的，一会儿运行任务A一会儿执行任务B，系统会不停地在两者间切换。但对于外部观察者来说，即使多个任务之间是串行并发的，也会造成多任务间是并行执行的错觉。

这两种情况在生活中都很常见。我曾经去黄山旅游过两次，黄山风景奇特，有着“五岳归来不看山，黄山归来不看岳”的美称。只要去过黄山的人都应该知道，导游时常挂在嘴边的“走路不看景，看景不走路”。因为黄山顶上经常下雨，地面湿滑，地形险峻。如果边走边看，跌倒擦伤那是常有的事。安全起见，就要求旅游在看景的时候，能够停下脚步，走路的时候能够专心看着地面，管好双脚。这就是“并发”。它和“边走边看”有着非常奇妙的关系，因为这两种情况，都可以被认为是“同时在看景和走路”。

那么在黄山上真正的“并行”应该是什么样子呢？聪明的同学应该可以想到，那就是坐缆车上山。缆车可以代替步行，你坐在缆车上才能专心欣赏沿途的风景，“走路”这些事情全部交给缆车去完成就好了。

实际上，如果系统内只有一个CPU，而使用多进程或者多线程任务，那么真实环境中这些任务不可能是真正并行的，毕竟一个CPU一次只能执行一条指令，这种情况下多进程或者多线程就是并发的，而不是并行的（操作系统会不停切换多个任务）。真正的并行也只可能出现在拥有多个CPU的系统中（比如多核CPU）。

由于并发的最终效果可能是和并行一样的，因此，如果没有特别的需要，我在本书中不会特别强调两者的区别。

1.2.3 临界区

临界区用来表示一种公共资源或者说是共享数据，可以被多个线程使用。但是每一次，只能有一个线程使用它，一旦临界区资源被占用，其他线程要想使用这个资源，就必须等待。

比如，在一个办公室里有一台打印机。打印机一次只能执行一个任务。如果小王和小明同时需要打印文件，很显然，如果小王先下发了打印任务，打印机就开始打印小王的文件。小明的任务就只能等待小王打印结束后才能打印。这里的打印机就是一个临界区的例子。

在并行程序中，临界区资源是保护的对象，如果意外出现打印机同时执行两个打印任务，那么最可能的结果就是打印出来的文件就会是损坏的文件。它既不是小王想要的，也不是小明想要的。

1.2.4 阻塞（**Blocking**）和非阻塞（**Non-Blocking**）

阻塞和非阻塞通常用来形容多线程间的相互影响。比如一个线程占用了临界区资源，那么其他所有需要这个资源的线程就必须在这个临界区中进行等待。等待会导致线程挂起，这种情况就是阻塞。此时，如果占用资源的线程一直不愿意释放资源，那么其他所有阻塞在这个临界区上的线程都不能工作。

非阻塞的意思与之相反，它强调没有一个线程可以妨碍其他线程执行。所有的线程都会尝试不断向前执行。有关这个概念，将在本章“并发级别”一节中做更详细的描述。

1.2.5 死锁（**Deadlock**）、饥饿（**Starvation**）和活锁（**Livelock**）

死锁、饥饿和活锁都属于多线程的活跃性问题。如果发现上述几种情况，那么相关线程可能就不再活跃，也就说它可能很难再继续往下执行了。

死锁应该是最糟糕的一种情况了（当然，其他几种情况也好不到哪里去），图1.6显示了一个死锁的发生。

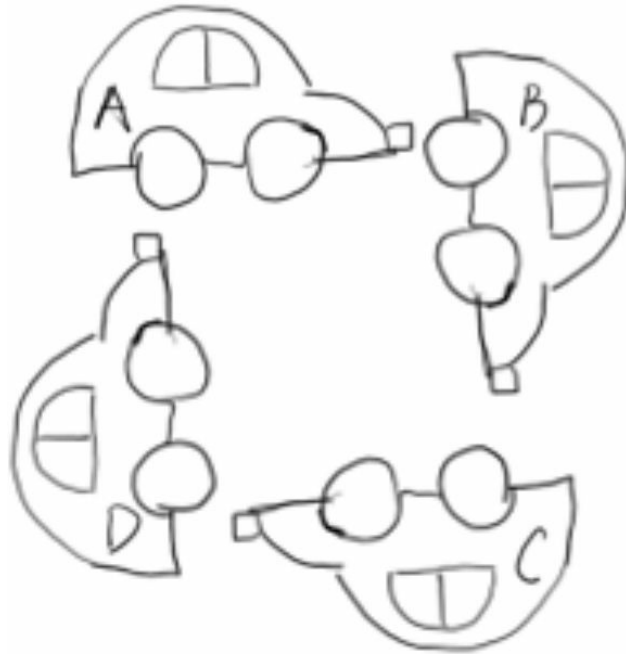


图1.6 死锁的发生

A、B、C、D四辆小车在这种情况下都无法继续行驶了。它们彼此之间相互占用了其他车辆的车道，如果大家都不愿意释放自己的车道，那么这个状态将永远维持下去，谁都不可能通过。死锁是一个很严重的，并且应该避免和时时小心的问题，我们将安排在“锁的优化与注意事项”一章中进行更详细的讨论。

饥饿是指某一个或者多个线程因为种种原因无法获得所需要的资源，导致一直无法执行。比如它的线程优先级可能太低，而高优先级的线程不断抢占它需要的资源，导致低优先级线程无法工作。在自然界中，母鸟喂食雏鸟时，很容易出现这种情况。由于雏鸟很多，食物可能有限，雏鸟之间的食物竞争可能非常厉害，小雏鸟因为经常抢不到食物，有可能会被饿死。线程的饥饿也非常类似这种情况。另外一种可能是，某一个线程一直占着关键资源不放，导致其他需要这个资源的线程无法正常执行，这种情况也是饥饿的一种。与死锁相比，饥饿还是有可能在未来一段时间内解决的（比如高优先级的线程已经完成任务，不再

疯狂的执行)。

活锁是一种非常有趣的情况。不知道大家是不是有遇到过这么一种场景，当你要坐电梯下楼，电梯到了，门开了，这时你正准备出去。但很不巧的是，门外一个人挡着你的去路，他想进来。于是，你很绅士地靠左走，避让对方。同时，对方也是非常绅士地，但他靠右走希望避让。结果，你们俩就又撞上了。于是乎，你们都意识到了问题，希望尽快避让对方，你立即向右边走，同时，他立即向左边走。结果，又撞上了！不过介于人类的智能，我相信这个动作重复2、3次后，你应该可以顺利解决这个问题。因为这个时候，大家都会本能的对视，进行交流，保证这种情况不再发生。

但如果这种情况发生在两个线程间可能就不会那么幸运了。如果线程的智力不够，且都秉承着“谦让”的原则，主动将资源释放给他人使用，那么就会出现资源不断在两个线程中跳动，而没有一个线程可以同时拿到所有资源而正常执行。这种情况就是活锁。

1.3 并发级别

由于临界区的存在，多线程之间的并发必须受到控制。根据控制并发的策略，我们可以把并发的级别进行分类，大致上可以分为阻塞、无饥饿、无障碍、无锁、无等待几种。

1.3.1 阻塞（**Blocking**）

一个线程是阻塞的，那么在其他线程释放资源之前，当前线程无法继续执行。当我们使用synchronized关键字，或者重入锁时（我们将在第2、3章介绍这两种技术），我们得到的就是阻塞的线程。

无论是synchronized或者重入锁，都会试图在执行后续代码前，得到临界区的锁，如果得不到，线程就会被挂起等待，直到占有了所需资源为止。

1.3.2 无饥饿（**Starvation-Free**）

如果线程之间是有优先级的，那么线程调度的时候总是会倾向于满足高优先级的线程。也就是说，对于同一个资源的分配，是不公平的！如图1.7所示，显示了非公平与公平两种情况（五角星表示高优先级线程）。对于非公平的锁来说，系统允许高优先级的线程插队。这样有可能导致低优先级线程产生饥饿。但如果锁是公平的，满足先来后到，那么饥饿就不会产生，不管新来的线程优先级多高，要想获得资源，就必

须乖乖排队。那么所有的线程都有机会执行。

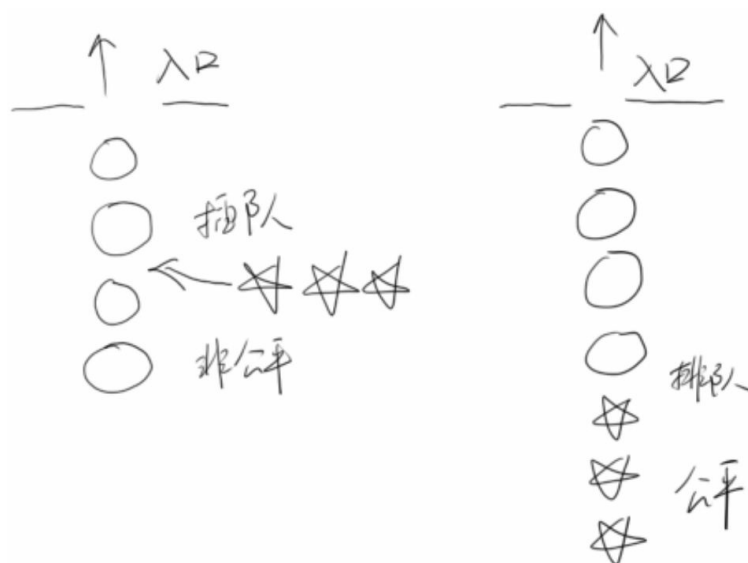


图1.7 公平与非公平锁

1.3.3 无障碍（Obstruction-Free）

无障碍是一种最弱的非阻塞调度。两个线程如果是无障碍的执行，那么他们不会因为临界区的问题导致一方被挂起。换言之，大家都可以大摇大摆地进入临界区了。那么如果大家一起修改共享数据，把数据改坏了可怎么办呢？对于无障碍的线程来说，一旦检测到这种情况，它就会立即对自己所做的修改进行回滚，确保数据安全。但如果没有数据竞争发生，那么线程就可以顺利完成自己的工作，走出临界区。

如果说阻塞的控制方式是悲观策略。也就是说，系统认为两个线程之间很有可能发生不幸的冲突，因此，以保护共享数据为第一优先级。相对来说，非阻塞的调度就是一种乐观的策略。它认为多个线程之间很有可能不会发生冲突，或者说这种概率不大。因此大家都应该无障碍的执行，但是一旦检测到冲突，就应该进行回滚。

从这个策略中也可以看到，无障碍的多线程程序并不一定能顺畅的运行。因为当临界区中存在严重的冲突时，所有的线程可能都会不断地回滚自己的操作，而没有一个线程可以走出临界区。这种情况会影响系统的正常执行。所以，我们可能会非常希望在这一堆线程中，至少可以有一个线程能够在有限的时间内完成自己的操作，而退出临界区。至少这样可以保证系统不会在临界区中进行无限的等待。

一种可行的无障碍实现可以依赖一个“一致性标记”来实现。线程在操作之前，先读取并保存这个标记，在操作完成后，再次读取，检查这个标记是否被更改过，如果两者是一致的，则说明资源访问没有冲突。如果不一致，则说明资源可能在操作过程中与其他写线程冲突，需要重试操作。而任何对资源有修改操作的线程，在修改数据前，都需要更新这个一致性标记，表示数据不再安全。

1.3.4 无锁（Lock-Free）

无锁的并行都是无障碍的。在无锁的情况下，所有的线程都能尝试对临界区进行访问，但不同的是，无锁的并发保证必然有一个线程能够在有限步内完成操作离开临界区。

在无锁的调用中，一个典型的特点是可能会包含一个无穷循环。在这个循环中，线程会不断尝试修改共享变量。如果没有冲突，修改成功，那么程序退出，否则继续尝试修改。但无论如何，无锁的并行总能保证有一个线程是可以胜出的，不至于全军覆没。至于临界区中竞争失败的线程，它们则必须不断重试，直到自己获胜。如果运气很不好，总是尝试不成功，则会出现类似饥饿的现象，线程会停止不前。

下面就是一段无锁的示意代码，如果修改不成功，那么循环永远不会停止。

```
while (!atomicVar.compareAndSet(localVar, localVar+1)) {  
    localVar = atomicVar.get();  
}
```

有关无锁，我们将安排在“锁的优化与注意事项”一章中详细介绍。

1.3.5 无等待（**Wait-Free**）

无锁只要求有一个线程可以在有限步内完成操作，而无等待则在无锁的基础上更进一步进行扩展。它要求所有的线程都必须在有限步内完成，这样就不会引起饥饿问题。如果限制这个步骤上限，还可以进一步分解为有界无等待和线程数无关的无等待几种，它们之间的区别只是对循环次数的限制不同。

一种典型的无等待结构就是RCU（Read-Copy-Update）。它的基本思想是，对数据的读可以不加控制。因此，所有的读线程都是无等待的，它们既不会被锁定等待也不会引起任何冲突。但在写数据的时候，先取得原始数据的副本，接着只修改副本数据（这就是为什么读可以不加控制），修改完成后，在合适的时机回写数据。

1.4 有关并行的两个重要定律

有关为什么要使用并程序的问题在之前已经进行了简单的探讨。总的来说，最重要的应该是出于两个目的。第一，为了获得更好的性能；第二，由于业务模型的需要，确实需要多个执行实体。在这里，我将更加关注于第一种情况，也就是有关性能的问题。将串程序改造为并发，一般来说可以提供程序的整体性能，但是究竟能提高多少，甚至说究竟是否真的可以提高，还是一个需要研究的问题。目前，主要有两个定律对这个问题进行解答，一个是Amdahl定律，另外一个Gustafson定律。

1.4.1 Amdahl定律

Amdahl定律是计算机科学中非常重要的定律。它定义了串行系统并行化后的加速比的计算公式和理论上限。

加速比定义：加速比 = 优化前系统耗时 / 优化后系统耗时

即，所谓加速比，就是优化前的耗时与优化后耗时的比值。加速比越高，表明优化效果越明显。图1.8显示了Amdahl公式的推导过程，其中 n 表示处理器个数， T 表示时间， T_1 表示优化前耗时（也就是只有1个处理器时的耗时）， T_n 表示使用 n 个处理器优化后的耗时。 F 是程序中只能串行执行的比例。

$$T_n = T_1 \left(F + \frac{1}{n} (1-F) \right)$$

\downarrow 处理器数 \nwarrow 串行比例 \searrow 并行比例 \swarrow 代入

$$\text{加速比} = \frac{T_1 \rightarrow \text{优化前耗时}}{T_n \rightarrow \text{优化后耗时}}$$

$$= \frac{T_1}{T_1 \left(F + \frac{1}{n} (1-F) \right)}$$

$$= \frac{1}{F + \frac{1}{n} (1-F)}$$

图1.8 Amdahl公式的推导

根据这个公式，如果CPU处理器数量趋于无穷，那么加速比与系统的串行化率成反比，如果系统中必须有50%的代码串行执行，那么系统的最大加速比为2。

假设有一程序分为以下步骤执行，每个执行步骤花费100个时间单位。其中，只有步骤2和步骤5可以进行并行，步骤1、3、4必须串行，如图1.9所示。在全串行的情况下，系统合计耗时500个时间单位。

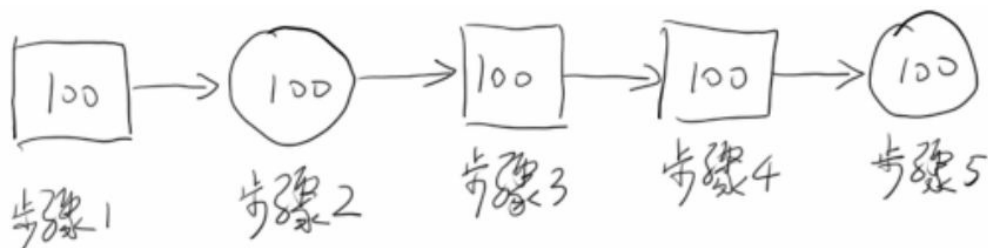


图1.9 串行工作流程

若将步骤2和步骤5并行化，假设在双核处理上，则有如图1.10所示的处理流程。在这种情况下，步骤2和步骤5的耗时将为50个时间单位。故系统整体耗时为400个时间单位。根据加速比的定义有：

$$\text{加速比} = \text{优化前系统耗时} / \text{优化后系统耗时} = 500 / 400 = 1.25$$

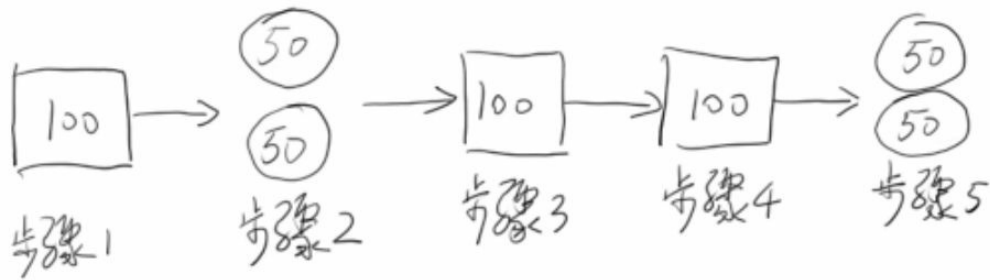


图1.10 双核处理上的并行化

或者根据前文中给出的加速比公式。由于5个步骤中，3个步骤必须串行，因此其串行化比重为 $3/5=0.6$ ，即 $F=0.6$ ，且双核处理器的处理器个数 N 为2。代入公式得：

$$\text{加速比} = 1 / (0.6 + (1 - 0.6) / 2) = 1.25$$

在极端情况下，假设并行处理器个数为无穷大，则有如图1.11所示的处理过程。步骤2和步骤5的处理时间趋于0。即使这样，系统整体耗时依然大于300个时间单位。即加速比的极限为 $500/300=1.67$ 。

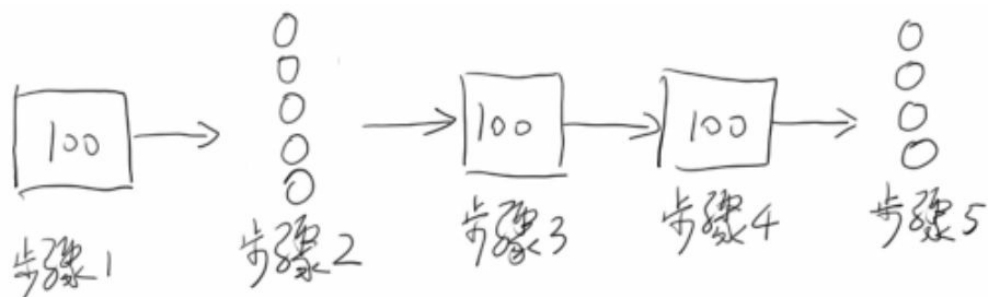


图1.11 极端情况下的并行化

使用加速比计算公式， N 趋于无穷大，有加速比 $=1/F$ ，且 $F=0.6$ ，故有加速比 $=1.67$ 。

由此可见，为了提高系统的速度，仅增加CPU处理器的数量并不一定能起到有效的作用。需要从根本上修改程序的串行行为，提高系统内可并行化的模块比重，在此基础上，合理增加并行处理器数量，才能以最小的投入，得到最大的加速比。

注意：根据Amdahl定律，使用多核CPU对系统进行优化，优化的效果取决于CPU的数量以及系统中的串行化程序的比重。CPU数量越多，串行化比重越低，则优化效果越好。仅提高CPU数量而不降低程序的串行化比重，也无法提高系统性能。

1.4.2 Gustafson定律

Gustafson定律也试图说明处理器个数、串行比例和加速比之间的关系，如图1.12所示，但是Gustafson定律和Amdahl定律的角度不同。同样，加速比都定义为优化前的系统耗时除以优化后的系统耗时。

$\xrightarrow{\text{串行时间}}$
 执行时间: $a+b$
 $\xrightarrow{\text{并行时间}}$
 总执行时间: $a+n \cdot b$
 $\xrightarrow{\text{处理器个数}}$

$$\text{加速比} = (a+n \cdot b) / (a+b)$$

定义: $F = a / (a+b)$ 串行比例

$$\begin{aligned}
 \text{则加速比 } S(n) &= \frac{a+n \cdot b}{a+b} = \frac{a}{a+b} + \frac{n \cdot b}{a+b} \\
 &= F + n \cdot \left(\frac{a+b-a}{a+b} \right) = F + n \left(1 - \frac{a}{a+b} \right) \\
 &= F + n(1-F) = F + n - nF \\
 &= n - F(n-1)
 \end{aligned}$$

图1.12 Gustafson定律的推导

可以看到，由于切入角度的不同，Gustafson定律的公式和Amdahl定律的公式截然不同。从Gustafson定律中，我们可以更容易地发现，如果串行化比例很小，并行化比例很大，那么加速比就是处理器的个数。只要你不断地累加处理器，就能获得更快的速度。

1.4.3 Amdahl定律和Gustafson定律是否相互矛盾

由于Amdahl定律和Gustafson定律的结论不同，这是不是说明这两个理论之间有一个是错误的呢？其实不然，两者的差异其实是因为这两个定律对同一个客观事实从不同角度去审视后的结果，它们的偏重点有

所不同。

举一个生活的例子，一辆汽车行驶在相聚60公里的城市。你花了一个小时，行驶了30公里。无论接下来开多快，你都不可能达到90公里/小时的时速。图1.13很好地说明了原因。

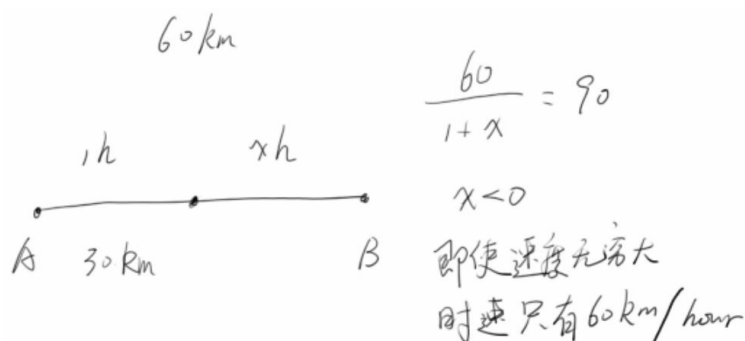


图1.13 Amdahl定律的偏重点

求解图1.13中的方程，你会发现如果你想达到90公里的时速，那么 你从AB中点到达B点的时间会是一个负数，这显然不是一个合理的结论。实际上，如果前半程30km你使用了一小时，那么即使你从中点到B点使用光速，也只能把整体的平均时速维持在60km/hour。

也就是说Amdahl强调：当串行比例一定时，加速比是有上限的，不管你堆叠多少个CPU参与计算，都不能突破这个上限！

而Gustafson定律的出发点与之不同，对Gustafson定律来说，不管 你从A点出发的速度有多慢，只要给你足够的时间和距离，只要你后期的速度比期望值快那么一点点，你总是可以把平均速度调整到非常接近那个期望值的。比如，你想要达到均速90km/hour，即使在前30km你的时速只有30km/hour，你只要在很后面的速度达到91km/hour，给你足够的时间和距离，你总有一天可以把均速提高到90km/hour。

因此，Gustafson定律关心的是：如果可被并行化的代码所占比重足

够多，那么加速比就能随着CPU的数量线性增长。

所以，这两个定律并不矛盾。从极端角度来说，如果系统中没有可被串行化的代码（即 $F=1$ ），那么对于这两个定律，其加速比都是1。反之，如果系统中可串行化代码比重达到100%，那么这两个定律得到加速比都是 n （处理器个数）。

1.5 回到Java: JMM

前面我已经介绍了有关并程序的一些关键概念和定律。这些概念可以说是与语言无关的。无论你使用Java或者C，或者其他任何一门语言编写并发程序，都有可能会涉及这些问题。但本书依然是一本面向Java程序员的书籍。因此，在本章最后，我们还是希望可以探讨一下有关Java的内存模型（JMM）。

由于并发程序要比串程序复杂很多，其中一个重要原因是并发程序下数据访问的一致性和安全性将会受到严重挑战。如何保证一个线程可以看到正确的数据呢？这个问题看起来很白痴。对于串程序来说，根本就是小菜一碟，如果你读取一个变量，这个变量的值是1，那么你读到的一定是1，就这么简单的问题在并程序中居然变得复杂起来。事实上，如果不加控制地任由线程胡乱并行，即使原本是1的数值，你也有可能读到2。因此，我们需要在深入了解并行机制的前提下，再定义一种规则，保证多个线程间可以有效地、正确地协同工作。而JMM也就是为此而生的。

JMM的关键技术点都是围绕着多线程的原子性、可见性和有序性来建立的。因此，我们首先必须了解这些概念。

1.5.1 原子性（Atomicity）

原子性是指一个操作是不可中断的。即使是在多个线程一起执行的时候，一个操作一旦开始，就不会被其他线程干扰。

比如，对于一个静态全局变量`int i`，两个线程同时对它赋值，线程A给他赋值1，线程B给它赋值为-1。那么不管这2个线程以何种方式、何种步调工作，`i`的值要么是1，要么是-1。线程A和线程B之间是没有干扰的。这就是原子性的一个特点，不可被中断。

但如果我们不使用`int`型而使用`long`型的话，可能就没有那么幸运了。对于32位系统来说，`long`型数据的读写不是原子性的（因为`long`有64位）。也就是说，如果两个线程同时对`long`进行写入的话（或者读取），对线程之间的结果是有干扰的。

大家可以仔细观察一下下面的代码：

```
public class MultiThreadLong {
    public static long t=0;
    public static class ChangeT implements Runnable{
        private long to;
        public ChangeT(long to){
            this.to=to;
        }
        @Override
        public void run() {
            while(true){
                MultiThreadLong.t=to;
                Thread.yield();
            }
        }
    }
    public static class ReadT implements Runnable{
```

```

@Override
public void run() {
    while(true){
        long tmp=MultiThreadLong.t;
        if(tmp!=111L && tmp!=-999L && tmp!=333L && tmp!=-444L)
            System.out.println(tmp);
        Thread.yield();
    }
}

public static void main(String[] args) {
    new Thread(new ChangeT(111L)).start();
    new Thread(new ChangeT(-999L)).start();
    new Thread(new ChangeT(333L)).start();
    new Thread(new ChangeT(-444L)).start();
    new Thread(new ReadT()).start();
}
}

```

上述代码有4个线程对long型数据t进行赋值，分别对t赋值为111、-999、333、444。然后，有一个读取线程，读取这个t的值。一般来说，t的值总是这4个数值中的一个。这当然也是我们的期望了。但很不幸，在32位的Java虚拟机中，未必总是这样。

如果读取线程ReadT总是读到合理的数据，那么这个程序应该没有任何输出。但是，实际上，这个程序一旦运行，就会大量输出以下信

息：（再次强调，使用32位虚拟机）

```
*****
-4294966963
4294966852
-4294966963
*****
```

这里截取了部分输出。我们可以看到，读取线程居然读到了两个似乎根本不可能存在的数值。这不是幻觉，在这里，你看到的确实是事实，其中的原因也就是因为32位系统中long型数据的读和写都不是原子性的，多线程之间相互干扰了！

如果我给出这几个数值的2进制表示，大家就会有更清晰的认识了：

[illegible]

上面显示了这几个相关数字的补码形式，也就是在计算机内的真实存储内容。不难发现，这个奇怪的4294966852，其实是111或者333的前32位，与-444的后32位夹杂后的数字。而-4294967185只是-999或者-444的前32位与111夹杂后的数字。换句话说，由于并行的关系，数字被写乱了，或者读的时候，读串位了。

通过这个例子，我想大家都原子性应该有了基本的认识。

1.5.2 可见性（**Visibility**）

可见性是指当一个线程修改了某一个共享变量的值，其他线程是否能够立即知道这个修改。显然，对于串行程序来说，可见性问题是存在的。因为你在任何一个操作步骤中修改了某个变量，那么在后续的步骤中，读取这个变量的值，一定是修改后的新值。

但是这个问题在并行程序中就不见得了。如果一个线程修改了某一个全局变量，那么其他线程未必可以马上知道这个改动。图1.14展示了发生可见性问题的一种可能。如果在CPU1和CPU2上各运行了一个线程，它们共享变量t，由于编译器优化或者硬件优化的缘故，在CPU1上的线程将变量t进行了优化，将其缓存在cache中或者寄存器里。这种情况下，如果在CPU2上的某个线程修改了变量t的实际值，那么CPU1上的线程可能并无法意识到这个改动，依然会读取cache中或者寄存器里的数据。因此，就产生了可见性问题。外在表现为：变量t的值被修改，但是CPU1上的线程依然会读到一个旧值。可见性问题也是并行程序开发中需要重点关注的问题之一。

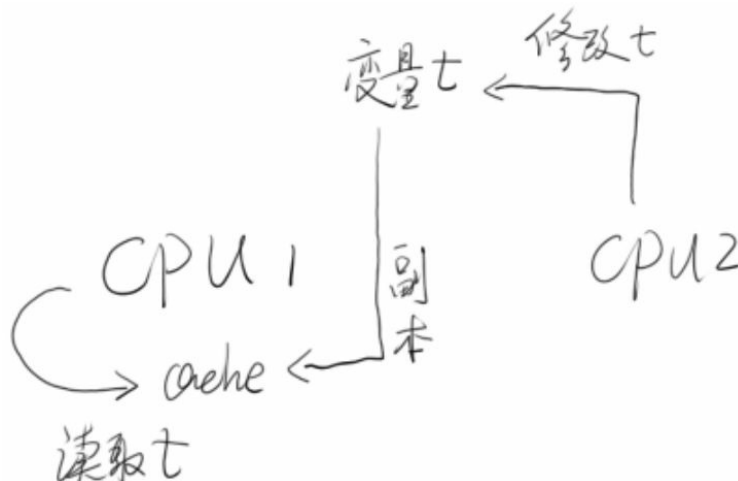


图1.14 可见性问题

可见性问题是一个综合性问题。除了上述提到的缓存优化或者硬件优化（有些内存读写可能不会立即触发，而会先进入一个硬件队列等待）会导致可见性问题外，指令重排（这个问题将在下一节中更详细讨论）以及编辑器的优化，都有可能导致一个线程的修改不会立即被其他线程察觉。

下面来看一个简单的例子：

Thread 1	Thread 2
1: r2 = A;	3: r1 = B;
2: B = 1;	4: A = 2;

上述两个线程，并行执行，分别有1、2、3、4四条指令。其中指令1、2属于线程1，而指令3、4属于线程2。

从指令的执行顺序上看， $r2==2$ 并且 $r1==1$ 似乎是不可能出现的。但实际上，我们并没有办法从理论上保证这种情况不出现。因为编译器可能将指令重排成：

Thread 1	Thread 2
B = 1;	r1 = B;
r2 = A;	A = 2;

在这种执行顺序中，就有可能出现刚才看似不可能出现的`r2==2`并且`r1==1`的情况了。

这个例子就说明，在一个线程中去观察另外一个线程的变量，它们的值是否能观测到、何时能观测到是没有保证的。

再来看一个稍微复杂一些的例子：

Thread 1	Thread 2
r1 = p;	r6 = p;
r2 = r1.x;	r6.x = 3;
r3 = q;	
r4 = r3.x;	
r5 = r1.x;	

这里假设在初始时，`p == q`并且`p.x == 0`。对于大部分编译器来说，可能会对线程1进行向前替换的优化，也就是`r5=r1.x`这条指令会被直接替换成`r5=r2`。因为它们都读取了`r1.x`，又发生在同一个线程中，因此，编译器很可能认为第2次读取是完全没有必要的。因此，上述指令可能会变成：

Thread 1	Thread 2
r1 = p;	r6 = p;
r2 = r1.x;	r6.x = 3;
r3 = q;	

```
r4 = r3.x;  
r5 = r2;
```

现在思考这么一种场景。假设线程2中的`r6.x=3`发生在`r2 = r1.x`和`r4 = r3.x`之间，而编译器又打算重用`r2`来表示`r5`。那么就有可能出现非常奇怪的现象。你看到的`r2`是0，`r4`是3，但是`r5`还是0。因此，如果从线程1代码的直观感觉上看就是：`p.x`的值从0变成了3（因为`r4`是3），接着又变成了0（这是不是算一个非常怪异的问题呢？）。

1.5.3 有序性（Ordering）

有序性问题可能是三个问题中最难理解的了。对于一个线程的执行代码而言，我们总是习惯地认为代码的执行是从先往后，依次执行的。这么理解也不能说完全错误，因为就一个线程内而言，确实会表现成这样。但是，在并发时，程序的执行可能就会出现乱序。给人直观的感觉就是：写在前面的代码，会在后面执行。听起来有些不可思议，是吗？有序性问题的原因是因为程序在执行时，可能会进行指令重排，重排后的指令与原指令的顺序未必一致。下面来看一个简单的例子：

```
01 class OrderExample {  
02     int a = 0;  
03     boolean flag = false;  
04     public void writer() {  
05         a = 1;  
06         flag = true;  
07     }  
08     public void reader() {
```

```
09     if (flag) {
10         int i = a + 1;
11         .....
12     }
13 }
14 }
```

假设线程A首先执行writer()方法，接着线程B执行reader()方法，如果发生指令重排，那么线程B在代码第10行时，不一定能看到a已经被赋值为1了。如图1.15所示，显示了两个线程的调用关系。

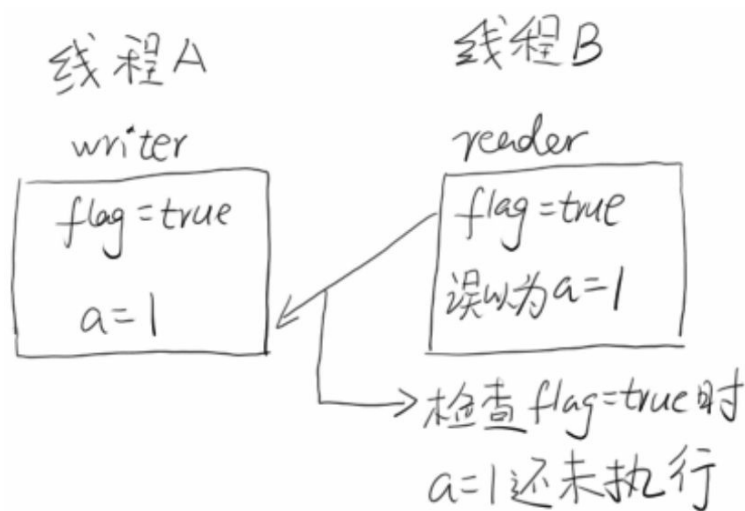


图1.15 指令重排引起线程间语义不一致

这确实是一个看起来很奇怪的问题，但是它确实可能存在。注意：我这里说的是可能存在。因为如果指令没有重排，这个问题就不存在了，但是指令是否发生重排、如何重排，恐怕是我们无法预测的。因此，对于这类问题，我认为比较严谨的描述是：线程A的指令执行顺序在线程B看来是没有保证的。如果运气好的话，线程B也许真的可以看到和线程A一样的执行顺序。

不过这里还需要强调一点，对于一个线程来说，它看到的指令执行顺序一定是一致的（否则的话我们的应用根本无法正常工作，不是吗？）。也就是说指令重排是有一个基本前提的，就是保证串行语义的一致性。指令重排不会使串行的语义逻辑发生问题。因此，在串行代码中，大可不必担心。

注意：指令重排可以保证串行语义一致，但是没有义务保证多线程间的语义也一致。

那么，好奇的你可能马上就会在脑海里闪出一个疑问，为什么要指令重排呢？让他一步一步执行多好呀！也不会有那么多奇葩的问题。

之所以那么做，完全是因为性能考虑。我们知道，一条指令的执行是可以分为很多步骤的。简单地说，可以分为以下几步：

- 取指IF
- 译码和取寄存器操作数ID
- 执行或者有效地址计算EX
- 存储器访问MEM
- 写回WB

我们的汇编指令也不是一步就可以执行完毕的，在CPU中实际工作时，它还是需要分为多个步骤依次执行的。当然，每个步骤所涉及的硬件也可能不同。比如，取指时会用到PC寄存器和存储器，译码时会用到指令寄存器组，执行时会使用ALU，写回时需要寄存器组。

注意：ALU指算术逻辑单元。它是CPU的执行单元，是CPU的核心组成部分，主要功能进行二进制算术运算。

由于每一个步骤都可能使用不同的硬件完成，因此，聪明的工程师们就发明了流水线技术来执行指令，如图1.16所示，显示了流水线的工作原理。

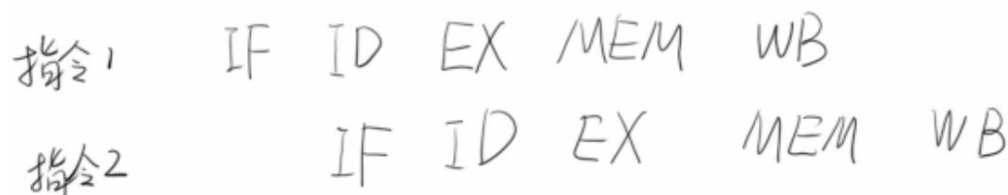


图1.16 指令流水线

可以看到，当第2条指令执行时，第1条执行其实并未执行完，确切地说第一条指令还没开始执行，只是刚刚完成了取值操作而已。这样的好处非常明显，假如这里每一个步骤都需要花费1毫秒，那么指令2等待指令1完全执行后，再执行，则需要等待5毫秒，而使用流水线后，指令2只需要等待1毫秒就可以执行了。如此大的性能提升，当然让人眼红。更何况，实际的商业CPU的流水线级别甚至可以达到10级以上，则性能提升就更加明显。

有了流水线这个神器，我们CPU才能真正高效的执行，但是，别忘了一点，流水线总是害怕被中断的。流水线满载时，性能确实相当不错，但是一旦中断，所有的硬件设备都会进入一个停顿期，再次满载又需要几个周期，因此，性能损失会比较大。所以，我们必须要想办法尽量不让流水线中断！

那么答案就来了，之所以需要做指令重排，就是为了尽量少的中断流水线。当然了，指令重排只是减少中断的一种技术，实际上，在CPU的设计中，我们还会使用更多的软硬件技术来防止中断，不过对它们的讨论已经远远超出本书范围，有兴趣的读者可以查阅相关资料。

让我们来仔细看一个例子。图1.17展示了 $A=B+C$ 这个操作的执行过程。写在左边的指令就是汇编指令。LW表示load，其中LW R1,B，表示把B的值加载到R1寄存器中。ADD指令就是加法，把R1、R2的值相加，并存放到R3中。SW表示store，存储，就是将R3寄存器的值保存到变量A中。

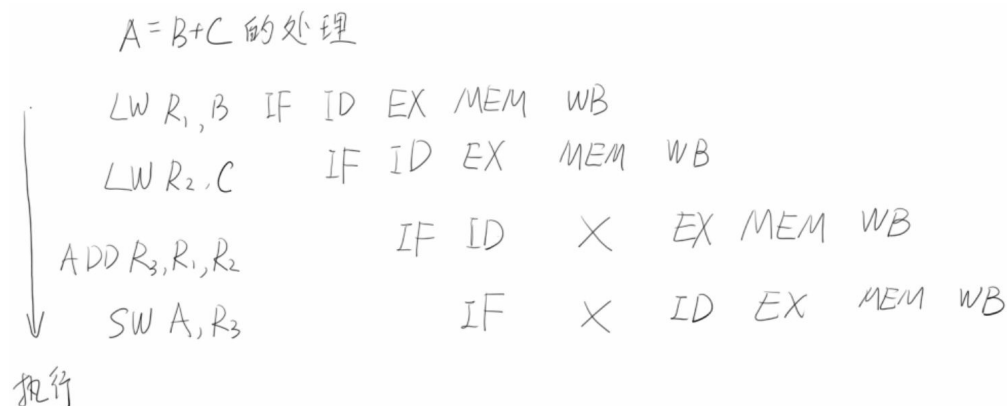


图1.17 $A=B+C$ 的执行过程

右边就是流水线的情况。注意，在ADD指令上，有一个大叉，表示一个中断。也就是说ADD在这里停顿了一下。为什么ADD会在这里停顿呢？原因很简单，R2中的数据还没有准备好！所以，ADD操作必须进行一次等待。由于ADD的延迟，导致其后面所有的指令都要慢一个节拍。

理解了上面这个例子，我们就可以来看一个更加复杂的情况：

```
a=b+c
d=e-f
```

上述代码的执行应该会是这样，如图1.18所示。

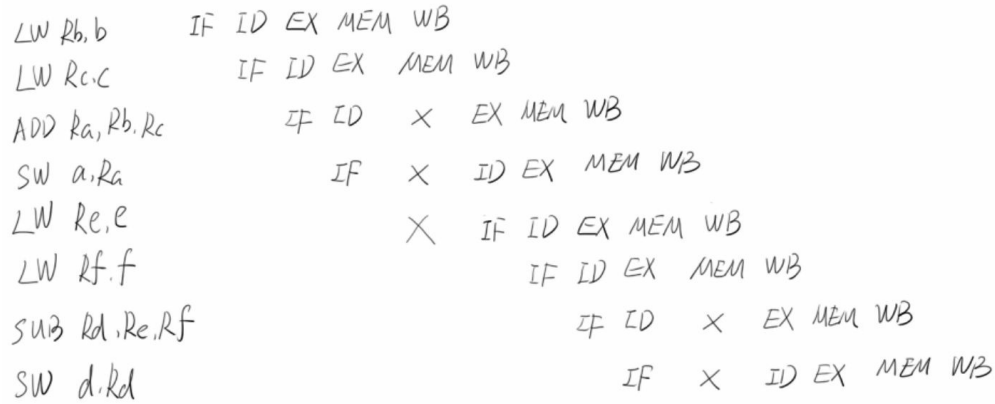


图1.18 重排前指令执行过程

由于ADD和SUB都需要等待上一条指令的结果，因此，在这里插入了不少停顿。那么对于这段代码，是否有可能消除这些停顿呢？显然是可以的，如图1.19所示，显示了减少这些停顿的方法。我们只需要将LW Re, e和LW Rf, f移动到前面执行即可。思想很简单，先加载e和f对程序是没有影响的。既然在ADD的时候一定要停顿一下，那么停顿的时间还不如去做点有意义的事情。

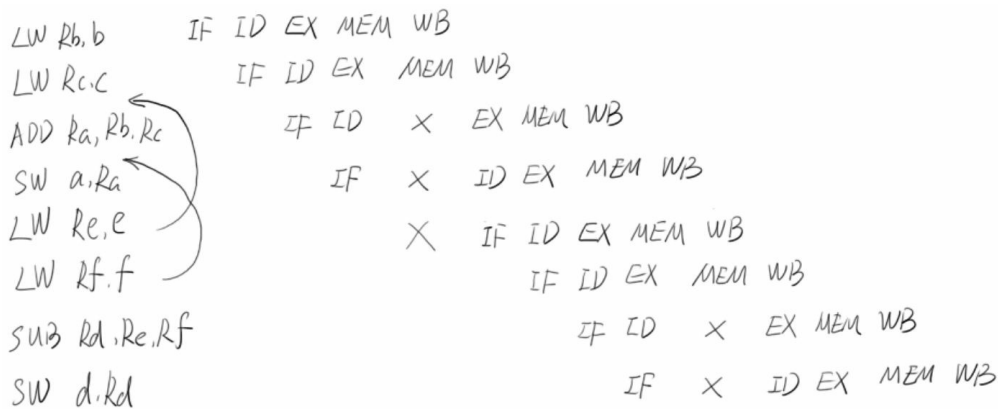


图1.19 指令重排，以消除停顿

重排后，最终的结果如图1.20所示。可以看到，所有的停顿都已经消除，流水线已经可以十分顺畅地执行。

LW Rb,b	IF ID EX MEM WB
LW Rc,c	IF ID EX MEM WB
LW Re,e	IF ID EX MEM WB
ADD Ra,Rb,Rc	IF ID EX MEM WB
LW Rf,f	IF ID EX MEM WB
SW a,Ra	IF ID EX MEM WB
SUB Rd,Re,Rf	IF ID EX MEM WB
SW d,Rd	IF ID EX MEM WB

图1.20 重排后的指令

由此可见，指令重排对于提高CPU处理性能是十分必要的。虽然确实带来了乱序的问题，但是这点牺牲是完全值得的。

1.5.4 哪些指令不能重排：Happen-Before规则

在前文已经介绍了指令重排，虽然Java虚拟机和执行系统会对指令进行一定的重排，但是指令重排是有原则的，并非所有的指令都可以随便改变执行位置，以下罗列了一些基本原则，这些原则是指令重排不可违背的。

- 程序顺序原则：一个线程内保证语义的串行性
- volatile规则：volatile变量的写，先发生于读，这保证了volatile变量的可见性
- 锁规则：解锁（unlock）必然发生在随后的加锁（lock）前
- 传递性：A先于B，B先于C，那么A必然先于C

- 线程的start()方法先于它的每一个动作
- 线程的所有操作先于线程的终结（Thread.join()）
- 线程的中断（interrupt()）先于被中断线程的代码
- 对象的构造函数执行、结束先于finalize()方法

以程序顺序原则为例，重排后的指令绝对不能改变原有的串行语义。比如：

```
a=1;  
b=a+1;
```

由于第2条语句依赖第一条的执行结果。如果冒然交换两条语句的执行顺序，那么程序的语义就会修改。因此这种情况是绝对不允许发生的。因此，这也是指令重排的一条基本原则。

此外，锁规则强调，unlock操作必然发生在后续的对同一个锁的lock之前。也就是说，如果对一个锁解锁后，再加锁，那么加锁的动作绝对不能重排到解锁动作之前。很显然，如果这么做，加锁行为是无法获得这把锁的。

其他几条原则也是类似的，这些原则都是为了保证指令重排不会破坏原有的语义结构。

1.6 参考文献

- Linus Torvalds: 忘掉那该死的并行吧!
 - <http://www.csdn.net/article/2015-01-08/2823487-linus-the-whole-parallel-computing-is-the-future-is-a-bunch>
 - <http://www.realworldtech.com/forum/?threadid=146066&curpostid=146227>
- 有关唐纳德
 - <http://blog.sciencenet.cn/blog-1225851-840243.html>
- 有关摩尔定律失效
 - <http://blog.csdn.net/hsutter/article/details/1136281>
 - <http://www.zdnet.com/article/barrett-still-has-some-fight-in-him>
- 有关并发的级别
 - <http://concurrencyfreaks.blogspot.hk/2013/05/lock-free-and-wait-free-definition-and.html>
 - <http://chuansong.me/n/862673>
- Amdahl定律
 - http://en.wikipedia.org/wiki/Amdahl's_law
- Gustafson定律

- http://en.wikipedia.org/wiki/Gustafson's_law
- 有关JMM
 - <http://docs.oracle.com/javase/specs/jls/se7/html/jls-17.html#jls-17.4>
- 有关指令重排
 - 《计算机体系结构》. 浙江大学出版社. 石教英等编

第2章 Java并行程序基础

我们已经探讨为什么必须面对并行程序这样复杂的程序设计方法，那么下面就需要静下心来，认真研究如何才能构建一个正确、健壮并且高效的并行系统。本章将详细介绍有关Java并行程序的设计基础，以及一些常见的问题，希望对读者有所帮助。

2.1 有关线程你必须知道的事

在介绍线程前，我们还是先了解一下线程的“母亲”——进程。如果你有读过操作系统的课程，那你对进程一定不会陌生。在这种专业级的书籍中，应该会给出一些“官方”的解释，比如像下面这样描述：

进程（**Process**）是计算机中的程序关于某数据集合上的一次运行活动，是系统进行资源分配和调度的基本单位，是操作系统结构的基础。在早期面向进程设计的计算机结构中，进程是程序的基本执行实体；在当代面向线程设计的计算机结构中，进程是线程的容器。程序是指令、数据及其组织形式的描述，进程是程序的实体。

不过我不想把这种严谨且抽象的描述介绍给大家。用一句简单的话来说，你在windows中，看到的后缀为.exe的文件，都是一个程序。不过程序是死的，静态的。当你双击这个.exe执行的时候，这个.exe文件中的指令就会被加载，那么你就能得到一个有关这个.exe程序的一个进程。进程是“活”的，或者说是正在被执行的。图2.1使用任务管理器，显示了当前系统中的进程。



图2.1 系统进程信息

进程中可以容纳若干个线程。它们并不是看不见、摸不着的，也可以使用工具看到它们，如图2.2所示。

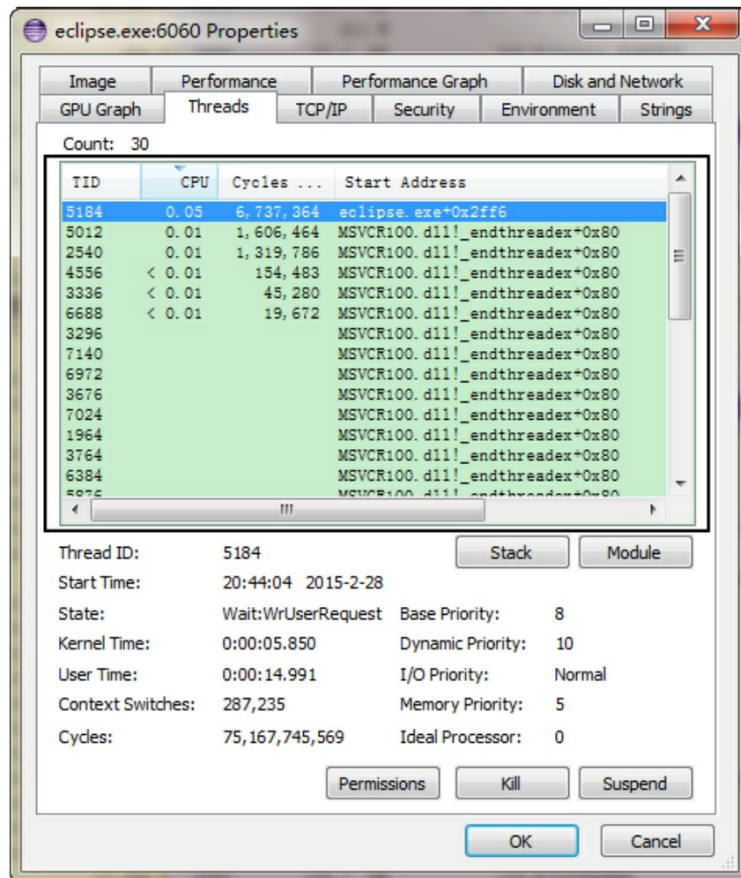


图2.2 进程中线程的信息

那线程和进程之间究竟是一种什么样的关系呢？简单地说，进程是一个容器。比如一间漂亮的小别墅。别墅里有电视、厨房、书房、洗手间等。当然，还有一家三口住在里面。当妈妈带女儿外出游玩时，爸爸一人在家。这时爸爸一个人在家里爱上哪里去哪里、爱干嘛干嘛，这时，爸爸就像一个线程（这个进程中只有一个活动线程），小别墅就像一个进程，家里的电视、厨房、书房就像这个进程占有的资源。当到三个人住在一起时（相当于三个线程），有时候可能就会有些小冲突，比如，当女儿占着电视机看动画片时，爸爸就不能看体育频道了，这就是线程间的资源竞争。当然，大部分时候，线程之间还是协作关系（如果我们创建线程是用来打架的，那创建它干嘛呢？）。比如，妈妈在厨房为爸爸和女儿做饭，爸爸在书房工作赚钱养家糊口，女儿在写作业，各

司其职，那么这个家就是其乐融融了，相对的，这个进程也就在健康地执行。

用稍微专业点的术语说，线程就是轻量级进程，是程序执行的最小单位。使用多线程而不是用多进程去进行并发程序的设计，是因为线程间的切换和调度的成本远远小于进程。

接下来让我们更细致地观察一个线程的生命周期。我们可以绘制一张简单的状态图来描述这个概念，如图2.3所示。

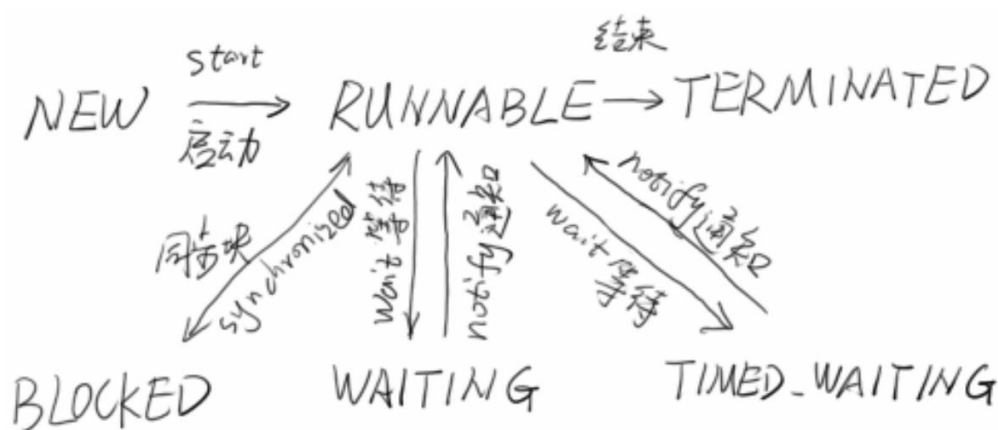


图2.3 线程状态图

线程的所有状态都在Thread中的State枚举中定义，如下所示：

```
public enum State {
    NEW,
    RUNNABLE,
    BLOCKED,
    WAITING,
    TIMED_WAITING,
    TERMINATED;
}
```

NEW状态表示刚刚创建的线程，这种线程还没开始执行。等到线程的start()方法调用时，才表示线程开始执行。当线程执行时，处于RUNNABLE状态，表示线程所需的一切资源都已经准备好了。如果线程在执行过程中遇到了synchronized同步块，就会进入BLOCKED阻塞状态，这时线程就会暂停执行，直到获得请求的锁。WAITING和TIMED_WAITING都表示等待状态，它们的区别是WAITING会进入一个无时间限制的等待，TIMED_WAITING会进行一个有时限的等待。那等待的线程究竟在等什么呢？一般来说，WAITING的线程正是在等待一些特殊的事件。比如，通过wait()方法等待的线程在等待notify()方法，而通过join()方法等待的线程则会等待目标线程的终止。一旦等到了期望的事件，线程就会再次执行，进入RUNNABLE状态。当线程执行完毕后，则进入TERMINATED状态，表示结束。

注意：从NEW状态出发后，线程不能再回到NEW状态，同理，处于TERMINATED的线程也不能再回到RUNNABLE状态。

2.2 初始线程：线程的基本操作

进行Java并发设计的第一步，就是必须要了解Java中为线程操作所提供的一些API。比如，如何新建并且启动线程，如何终止线程、中断线程等。当然了，因为并行操作要比串行操作复杂得多，于是，围绕着这些常用接口，可能有些比较隐晦的“坑”等着你去踩。而本节也将尽可能地将一些潜在问题描述清楚。

2.2.1 新建线程

新建线程很简单。只要使用new关键字创建一个线程对象，并且将它start()起来即可。

```
Thread t1=new Thread();  
t1.start();
```

那线程start()后，会干什么呢？这才是问题的关键。线程Thread，有一个run()方法，start()方法就会新建一个线程并让这个线程执行run()方法。

这里要注意，下面的代码也能通过编译，也能正常执行。但是，却不能新建一个线程，而是在当前线程中调用run()方法，只是作为一个普通的方法调用。

```
Thread t1=new Thread();  
t1.run();
```

因此，在这里希望大家特别注意，调用start()方法和直接调用run()方法的区别。

注意：不要用run()来开启新线程。它只会在当前线程中，串行执行run()中的代码。

默认情况下，Thread的run()方法什么都没有做，因此，这个线程一启动就马上结束了。如果你想让线程做点什么，就必须重载run()方法，把你的“任务”填进去。

```
Thread t1=new Thread(){
    @Override
    public void run(){
        System.out.println("Hello, I am t1");
    }
};
t1.start();
```

上述代码使用匿名内部类，重载了run()方法，并要求线程在执行时打印“Hello, I am t1”的字样。如果没有特别的需要，都可以通过继承Thread，重载run()方法来自定义线程。但考虑到Java是单继承的，也就是说继承本身也是一种很宝贵的资源，因此，我们也可以使用Runnable接口来实现同样的操作。Runnable接口是一个单方法接口，它只有一个run()方法：

```
public interface Runnable {
    public abstract void run();
}
```

此外，Thread类有一个非常重要的构造方法：

```
public Thread(Runnable target)
```

它传入一个Runnable接口的实例，在start()方法调用时，新的线程就会执行Runnable.run()方法。实际上，默认的Thread.run()就是这么做的：

```
public void run() {  
    if (target != null) {  
        target.run();  
    }  
}
```

注意：默认的Thread.run()就是直接调用内部的Runnable接口。因此，使用Runnable接口告诉线程该做什么，更为合理。

```
public class CreateThread3 implements Runnable {  
    public static void main(String[] args) {  
        Thread t1=new Thread(new CreateThread3());  
        t1.start();  
    }  
  
    @Override  
    public void run() {  
        System.out.println("Oh, I am Runnable");  
    }  
}
```

上述代码实现了Runnable接口，并将该实例传入Thread。这样避免重载Thread.run()，单纯使用接口来定义Thread，也是最常用的做法。

2.2.2 终止线程

一般来说，线程在执行完毕后就会结束，无须手工关闭。但是，凡事也都有例外。一些服务端的后台线程可能会常驻系统，它们通常不会正常终结。比如，它们的执行体本身就是一个大大的无穷循环，用于提供某些服务。

那如何正常的关闭一个线程呢？查阅JDK，你不难发现Thread提供了一个stop()方法。如果你使用stop()方法，就可以立即将一个线程终止，非常方便。但如果你使用的是eclipse之类的IDE写代码的话，就会立即发现stop()方法是一个被标注为废弃的方法。也就是说，在将来，JDK可能就会移除该方法。

为什么stop()被废弃而不推荐使用呢？原因是stop()方法太过于暴力，强行把执行到一半的线程终止，可能会引起一些数据不一致的问题。

为了让大家更好地理解本节内容，我先简单介绍一些有关数据不一致的概念。假设我们在数据库里维护着一张用户表，里面记录了用户ID和用户名。假设，这里有两条记录：

记录1: ID=1, NAME=小明

记录2: ID=2, NAME=小王

如果我们用一个User对象去保存这些记录，我们总是希望这个对象

要么保存记录1，要么保存记录2。如果这个User对象一半存着记录1，另外一半存在记录2，我想大部分人都会抓狂吧！如果现在真的由于程序问题，出现了这么一个怪异的对象u，u的ID是1，但是u的Name是小王。那么，我们说，在这种情况下，数据就已经不一致了。说白了就是系统有错误了。这种情况是相当危险的，如果我们把一个不一致的数据直接写入了数据库，那么就会造成数据永久地被破坏和丢失，后果不堪设想。

也许有人会问，怎么可能呢？跑得好好的系统，怎么会出这种问题呢？在单线程环境中，确实不会，但在并行程序中，如果考虑不周，就有可能出现类似的情况。不经思考地使用`stop()`就有可能导致这种问题。

`Thread.stop()`方法在结束线程时，会直接终止线程，并且会立即释放这个线程所持有的锁。而这些锁恰恰是用来维持对象一致性的。如果此时，写线程写入数据正写到一半，并强行终止，那么对象就会被写坏，同时，由于锁已经被释放，另外一个等待该锁的读线程就顺理成章的读到了这个不一致的对象，悲剧也就此发生。整个过程如图2.4所示。

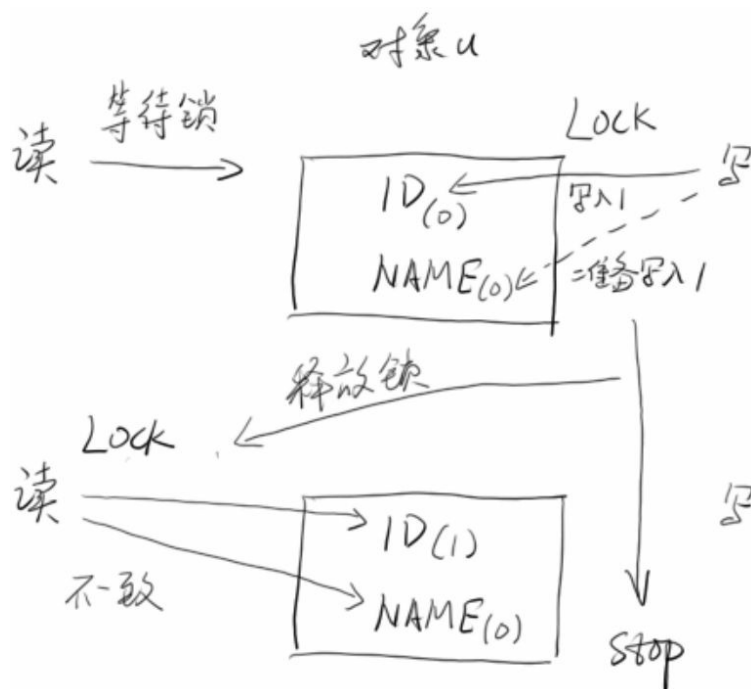


图2-4 stop() 方法强行终止线程导致数据不一致

首先，对象u持有ID和NAME两个字段，简单起见，这里假设当ID等于NAME时表示对象是一致的，否则表示对象出错。写线程总是会将ID和NAME写成相同的值，并且在这里初始值都为0。当写线程在写对象时，读线程由于无法获得锁，因此必须等待，所以读线程是看不见一个写了一半的对象的。当写线程写完ID后，很不幸地被stop()，此时对象u的ID为1而NAME仍然为0，处于不一致状态。而被终止的写线程简单地将锁释放，读线程争夺到锁后，读取数据，于是，读到了ID=1而NAME=0的错误值。

这个过程可以用以下代码模拟，这里读线程ReadObjectThread在读到对象的ID和NAME不一致时，会输出这些对象。而写线程ChangeObjectThread总是会写入两个相同的值。注意，代码在第56行会通过stop()方法强行终止写线程。

```
01 public class StopThreadUnsafe {
```

```
02     public static User u=new User();
03     public static class User{
04         private int id;
05         private String name;
06         public User(){
07             id=0;
08             name="0";
09         }
10         //省略setter和getter方法
11         @Override
12         public String toString() {
13             return "User [id=" + id + ", name=" + name + "]";
14         }
15     }
16     public static class ChangeObjectThread extends Thread{
17         @Override
18         public void run(){
19             while(true){
20                 synchronized(u){
21                     int v=(int)(System.currentTimeMillis())/100
22                     u.setId(v);
23                     //Oh, do sth. else
24                     try {
25                         Thread.sleep(100);
26                     } catch (InterruptedException e) {
27                         e.printStackTrace();
28                     }
29                 }
30             }
31         }
32     }
33 }
```

```
29             u.setName(String.valueOf(v));
30         }
31         Thread.yield();
32     }
33 }
34 }
35
36 public static class ReadObjectThread extends Thread{
37     @Override
38     public void run(){
39         while(true){
40             synchronized(u){
41                 if(u.getId() != Integer.parseInt(u.getName
42                     System.out.println(u.toString()));
43             }
44         }
45         Thread.yield();
46     }
47 }
48 }
49
50 public static void main(String[] args) throws InterruptedException
51     new ReadObjectThread().start();
52     while(true){
53         Thread t=new ChangeObjectThread();
54         t.start();
55         Thread.sleep(150);
```

```
56         t.stop();
57     }
58 }
59 }
```

执行以上代码，可以很容易得到类似如下输出，ID和NAME产生了一致。

```
User [id=1425135593, name=1425135592]
User [id=1425135594, name=1425135593]
```

如果在线上环境跑出以上结果，那么加班加点估计是免不了了，因为这类问题一旦出现，就很难排查，因为它们甚至没有任何错误信息，也没有线程堆栈。这种情况一旦混杂在动辄十几万行的程序代码中时，发现它们就全凭经验、时间还有一点点运气了。因此，除非你很清楚你在做什么，否则不要随便使用`stop()`方法来停止一个线程。

那如果需要停止一个线程时，应该这么做呢？其实方法很简单，只是需要由我们自行决定线程何时退出就可以了。仍然用本例说明，只需要将`ChangeObjectThread`线程增加一个`stopMe()`方法即可。如下所示：

```
01 public static class ChangeObjectThread extends Thread {
02     volatile boolean stopme = false;
03
04     public void stopMe(){
05         stopme = true;
06     }
07     @Override
```

```
08     public void run() {
09         while (true) {
10             if (stopme){
11                 System.out.println("exit by stop me");
12                 break;
13             }
14             synchronized (u) {
15                 int v = (int) (System.currentTimeMillis() / 10
16                 u.setId(v);
17                 //Oh, do sth. else
18                 try {
19                     Thread.sleep(100);
20                 } catch (InterruptedException e) {
21                     e.printStackTrace();
22                 }
23                 u.setName(String.valueOf(v));
24             }
25             Thread.yield();
26         }
27     }
28 }
```

代码第2行，定义了一个标记变量`stopme`，用于指示线程是否需要退出。当`stopMe()`方法被调用，`stopme`就被设置为`true`，此时，在代码第10行检测到这个改动时，线程就自然退出了。使用这种方式退出线程，不会使对象`u`的状态出现错误。因为，`ChangeObjectThread`已经不会有机会“写坏”对象了，它总是会选择一个合适的时间终止线程。

2.2.3 线程中断

在Java中，线程中断是一种重要的线程协作机制。从表面上理解，中断就是让目标线程停止执行的意思，实际上并非完全如此。在上一节中，我们已经详细讨论了`stop()`方法停止线程的害处，并且使用了一套自有的机制完善线程退出的功能。那在JDK中是否有提供更强大的支持呢？答案是肯定的，那就是线程中断。

严格地讲，线程中断并不会使线程立即退出，而是给线程发送一个通知，告知目标线程，有人希望你退出啦！至于目标线程接到通知后如何处理，则完全由目标线程自行决定。这点很重要，如果中断后，线程立即无条件退出，我们就又会遇到`stop()`方法的老问题。

与线程中断有关的，有三个方法，这三个方法看起来很像，所以可能会引起混淆和误用，希望大家注意。

```
public void Thread.interrupt()           // 中断线程
public boolean Thread.isInterrupted()     // 判断是否被中断
public static boolean Thread.interrupted() // 判断是否被中断，并清除
```

`Thread.interrupt()`方法是一个实例方法。它通知目标线程中断，也就是设置中断标志位。中断标志位表示当前线程已经被中断了。

`Thread.isInterrupted()`方法也是实例方法，它判断当前线程是否有被中断（通过检查中断标志位）。最后的静态方法`Thread.interrupted()`也是用来判断当前线程的中断状态，但同时会清除当前线程的中断标志位状态。

下面这段代码对t1线程进行了中断，那么中断后，t1会停止执行吗？

```

public static void main(String[] args) throws InterruptedException {
    Thread t1=new Thread(){
        @Override
        public void run(){
            while(true){
                Thread.yield();
            }
        }
    };
    t1.start();
    Thread.sleep(2000);
    t1.interrupt();
}

```

在这里，虽然对t1进行了中断，但是在t1中并没有中断处理的逻辑，因此，即使t1线程被置上了中断状态，但是这个中断不会发生任何作用。

如果希望t1在中断后退出，就必须为它增加相应的中断处理代码：

```

Thread t1=new Thread(){
    @Override
    public void run(){
        while(true){
            if(Thread.currentThread().isInterrupted()){
                System.out.println("Interruted!");
                break;
            }
        }
    }
}

```

```
        Thread.yield();
    }
}
};
```

上述代码的加粗部分使用`Thread.isInterrupted()`函数判断当前线程是否被中断了，如果是，则退出循环体，结束线程。这看起来与前面增加`stopme`标记的手法非常相似，但是中断的功能更为强劲。比如，如果在循环体中，出现了类似于`wait()`或者`sleep()`这样的操作，则只能通过中断来识别了。

下面，先来了解一下`Thread.sleep()`函数，它的签名如下：

```
public static native void sleep(long millis) throws InterruptedException
```

`Thread.sleep()`方法会让当前线程休眠若干时间，它会抛出一个`InterruptedException`中断异常。`InterruptedException`不是运行时异常，也就是说程序必须捕获并且处理它，当线程在`sleep()`休眠时，如果被中断，这个异常就会产生。

```
01 public static void main(String[] args) throws InterruptedException
02     Thread t1=new Thread(){
03         @Override
04         public void run(){
05             while(true){
06                 if(Thread.currentThread().isInterrupted()){
07                     System.out.println("Interruted!");
08                     break;
```



```
09         }
10         try {
11             Thread.sleep(2000);
12         } catch (InterruptedException e) {
13             System.out.println("Interruted When Sleep"
14                 //设置中断状态
15             Thread.currentThread().interrupt();
16         }
17         Thread.yield();
18     }
19 }
20 };
21 t1.start();
22 Thread.sleep(2000);
23 t1.interrupt();
24 }
```

注意上述代码中第10~15行加粗部分，如果在第11行代码处，线程被中断，则程序会抛出异常，并进入第13行处理。在catch子句部分，由于已经捕获了中断，我们可以立即退出线程。但在这里，我们并没有这么做，因为也许在这段代码中，我们还必须进行后续的处理，保证数据的一致性和完整性，因此，执行了Thread.interrupt()方法再次中断自己，置上中断标记位。只有这么做，在第6行的中断检查中，才能发现当前线程已经被中断了。

注意：Thread.sleep()方法由于中断而抛出异常，此时，它会清除中断标记，如果不加处理，那么在下次循环开始时，就无法捕获

这个中断，故在异常处理中，再次设置中断标记位。

2.2.4 等待（**wait**）和通知（**notify**）

为了支持多线程之间的协作，JDK提供了两个非常重要的接口线程等待`wait()`方法和通知`notify()`方法。这两个方法并不是在`Thread`类中的，而是输出`Object`类。这也意味着任何对象都可以调用这两个方法。

这两个方法的签名如下：

```
public final void wait() throws InterruptedException
public final native void notify()
```

当在一个对象实例上调用`wait()`方法后，当前线程就会在这个对象上等待。这是什么意思呢？比如，线程A中，调用了`obj.wait()`方法，那么线程A就会停止继续执行，而转为等待状态。等待到何时结束呢？线程A会一直等到其他线程调用了`obj.notify()`方法为止。这时，`obj`对象就俨然成为多个线程之间的有效通信手段。

那`wait()`和`notify()`究竟是如何工作的呢？图2.5展示了两者的工作过程。如果一个线程调用了`object.wait()`，那么它就会进入`object`对象的等待队列。这个等待队列中，可能会有多个线程，因为系统运行多个线程同时等待某一个对象。当`object.notify()`被调用时，它就会从这个等待队列中，随机选择一个线程，并将其唤醒。这里希望大家注意的是，这个选择是不公平的，并不是先等待的线程会优先被选择，这个选择完全是随机的。

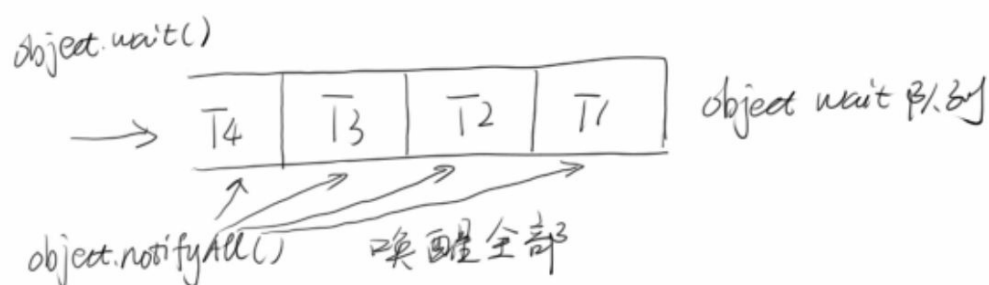
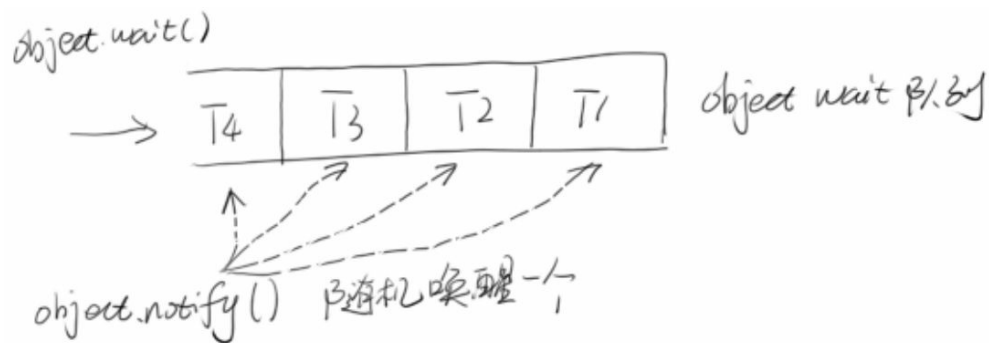


图2.5 notify() 唤醒等待的线程

除了notify()方法外，Object对象还有一个类似的notifyAll()方法，它和notify()的功能基本一致，但不同的是，它会唤醒在这个等待队列中所有等待的线程，而不是随机选择一个。

这里还需要强调一点，Object.wait()方法并不是可以随便调用的。它必须包含在对应的synchronized语句中，无论是wait()或者notify()都需要首先获得目标对象的一个监视器。如图2.6所示，显示了wait()和notify()的工作流程细节。其中T1和T2表示两个线程。T1在正确执行wait()方法前，首先必须获得object对象的监视器。而wait()方法在执行后，会释放这个监视器。这样做的目的是使得其他等待在object对象上的线程不至于因为T1的休眠而全部无法正常运行。

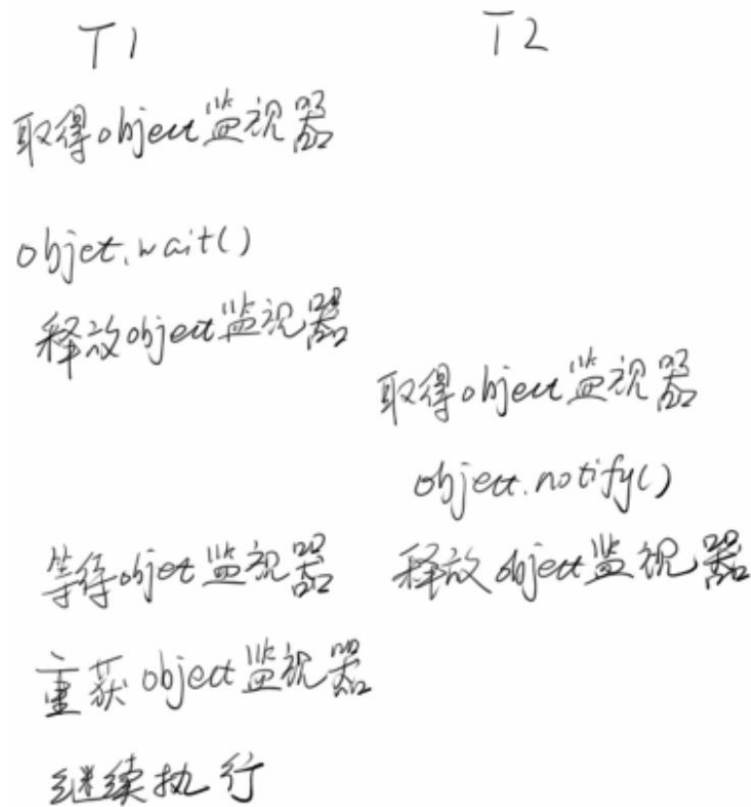


图2.6 wait()和notify()的工作流程细节

线程T2在notify()调用前，也必须获得object的监视器。所幸，此时T1已经释放了这个监视器。因此，T2可以顺利获得object的监视器。接着，T2执行了notify()方法尝试唤醒一个等待线程，这里假设唤醒了T1。T1在被唤醒后，要做的第一件事并不是执行后续的代码，而是要尝试重新获得object的监视器，而这个监视器也正是T1在wait()方法执行前所持有的那个。如果暂时无法获得，T1还必须要等待这个监视器。当监视器顺利获得后，T1才可以真正意义上的继续执行。

为了方便大家理解，这里给出一个简单地使用wait()和notify()的案例：

```
01 public class SimpleWN {  
02     final static Object object = new Object();
```

```
03     public static class T1 extends Thread{
04         public void run()
05         {
06             synchronized (object) {
07                 System.out.println(System.currentTimeMillis()+
08                 try {
09                     System.out.println(System.currentTimeMillis()
10                     object.wait();
11                 } catch (InterruptedException e) {
12                     e.printStackTrace();
13                 }
14                 System.out.println(System.currentTimeMillis()+
15             }
16         }
17     }
18     public static class T2 extends Thread{
19         public void run()
20         {
21             synchronized (object) {
22                 System.out.println(System.currentTimeMillis()+
23                 thread");
24                 object.notify();
25                 System.out.println(System.currentTimeMillis()+
26                 try {
27                     Thread.sleep(2000);
28                 } catch (InterruptedException e) {
29                 }
```

```

29         }
30     }
31 }
32 public static void main(String[] args) {
33     Thread t1 = new T1() ;
34     Thread t2 = new T2() ;
35     t1.start();
36     t2.start();
37 }
38 }

```

上述代码中，开启了两个线程T1和T2。T1执行了object.wait()方法。注意，在程序第6行，执行wait()方法前，T1先申请object的对象锁。因此，在执行object.wait()时，它是持有object的锁的。wait()方法执行后，T1会进行等待，并释放object的锁。T2在执行notify()之前也会先获得object的对象锁。这里为了让实验效果明显，特意安排在notify()执行之后，让T2休眠2秒钟，这样做可以更明显地说明，T1在得到notify()通知后，还是会先尝试重新获得object的对象锁。上述代码的执行结果类似如下：

```

1425224592258:T1 start!
1425224592258:T1 wait for object
1425224592258:T2 start! notify one thread
1425224592258:T2 end!
1425224594258:T1 end!

```

注意程序打印的时间戳信息，可以看到，在T2通知T1继续执行后，T1并不能立即继续执行，而是要等待T2释放object的锁，并重新成

功获得锁后，才能继续执行。因此，加粗部分时间戳的间隔为2秒（因为T2休眠了2秒）。

注意：`Object.wait()`和`Thread.sleep()`方法都可以让线程等待若干时间。除了`wait()`可以被唤醒外，另外一个主要区别就是`wait()`方法会释放目标对象的锁，而`Thread.sleep()`方法不会释放任何资源。

2.2.5 挂起（**suspend**）和继续执行（**resume**）线程

如果你阅读JDK有关Thread类的API文档，可能还会发现两个看起来非常有用的接口，即线程挂起（`suspend`）和继续执行（`resume`）。这两个操作是一对相反的操作，被挂起的线程，必须要等到`resume()`操作后，才能继续指定。乍看之下，这对操作就像`Thread.stop()`方法一样好用。但如果你仔细阅读文档说明，会发现它们也早已被标注为废弃方法，并不推荐使用。

不推荐使用`suspend()`去挂起线程的原因，是因为`suspend()`在导致线程暂停的同时，并不会去释放任何锁资源。此时，其他任何线程想要访问被它暂用的锁时，都会被牵连，导致无法正常继续运行（如图2.7所示）。直到对应的线程上进行了`resume()`操作，被挂起的线程才能继续，从而其他所有阻塞在相关锁上的线程也可以继续执行。但是，如果`resume()`操作意外地在`suspend()`前就执行了，那么被挂起的线程可能很难有机会被继续执行。并且，更严重的是：它所占用的锁不会被释放，因此可能会导致整个系统工作不正常。而且，对于被挂起的线程，从它

的线程状态上看，居然还是Runnable，这也会严重影响我们对系统当前状态的判断。

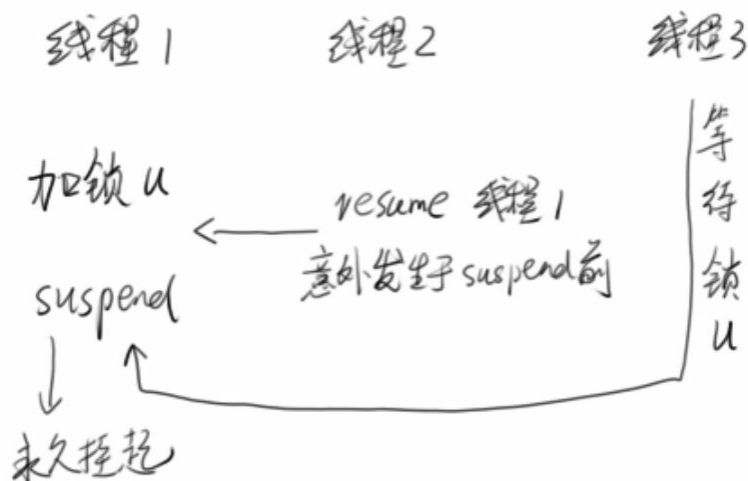


图2.7 suspend() 方法导致线程进入类似死锁的状态

为了方便大家理解suspend()的问题，这里准备一个简单的程序。演示了这种情况：

```
01 public class BadSuspend {
02     public static Object u = new Object();
03     static ChangeObjectThread t1 = new ChangeObjectThread("t1")
04     static ChangeObjectThread t2 = new ChangeObjectThread("t2")
05
06     public static class ChangeObjectThread extends Thread {
07         public ChangeObjectThread(String name){
08             super.setName(name);
09         }
10         @Override
11         public void run() {
12             synchronized (u) {
```



```

13         System.out.println("in "+getName());
14         Thread.currentThread().suspend();
15     }
16 }
17 }
18
19 public static void main(String[] args) throws InterruptedException
20     {
21         t1.start();
22         Thread.sleep(100);
23         t2.start();
24         t1.resume();
25         t2.resume();
26         t1.join();
27         t2.join();
28     }

```

执行上述代码，开启t1和t2两个线程。他们会在第12行通过对象锁u实现对临界区的访问。线程t1和t2启动后，在主函数中，第23~24行，对其进行resume()。目的是让他们得以继续执行。接着，主函数等待着两个线程的结束。

执行上述代码后，我们可能会得到以下输出：

```

in t1
in t2

```

这表明两个线程先后进入了临界区。但是程序不会退出。而是会挂

起。使用jstack命令打印系统的线程信息可以看到：

```
"t2" #9 prio=5 os_prio=0 tid=0x15c85c00 nid=0x1ddc runnable [0x15
  java.lang.Thread.State: RUNNABLE
    at java.lang.Thread.suspend0(Native Method)
    at java.lang.Thread.suspend(Thread.java:1029)
    at geym.conc.ch2.suspend.BadSuspend$ChangeObjectThread.ru
    - locked <0x048b2e58> (a java.lang.Object)
```

这时我们需要注意，当前系统中，线程t2其实是被挂起的。但是它的线程状态确实是**RUNNABLE**，这很有可能使我们误判当前系统的状态。同时，虽然主函数中已经调用了resume()，但是由于时间先后顺序的缘故，那个resume并没有生效！这就导致了线程t2被永远挂起，并且永远占用了对象u的锁。这对于系统来说极有可能是致命的。

如果需要一个比较可靠的suspend()函数，那应该怎么办呢？回想一下上一节中提到的wait()和notify()方法，这也不是一件难事。下面的代码就给出了一个利用wait()和notify()方法，在应用层面实现suspend()和resume()功能的例子。

```
01 public class GoodSuspend {
02     public static Object u = new Object();
03
04     public static class ChangeObjectThread extends Thread {
05         volatile boolean suspendme = false;
06
07         public void suspendMe() {
08             suspendme = true;
```

```
09         }
10
11     public void resumeMe(){
12         suspendme=false;
13         synchronized (this){
14             notify();
15         }
16     }
17     @Override
18     public void run() {
19         while (true) {
20
21             synchronized (this) {
22                 while (suspendme)
23                     try {
24                         wait();
25                     } catch (InterruptedException e) {
26                         e.printStackTrace();
27                     }
28             }
29
30             synchronized (u) {
31                 System.out.println("in ChangeObjectThread"
32             }
33             Thread.yield();
34         }
35     }
```

```
36     }
37
38     public static class ReadObjectThread extends Thread {
39         @Override
40         public void run() {
41             while (true) {
42                 synchronized (u) {
43                     System.out.println("in ReadObjectThread");
44                 }
45                 Thread.yield();
46             }
47         }
48     }
49
50     public static void main(String[] args) throws InterruptedException
51     {
52         ChangeObjectThread t1 = new ChangeObjectThread();
53         ReadObjectThread t2 = new ReadObjectThread();
54         t1.start();
55         t2.start();
56         Thread.sleep(1000);
57         t1.suspendMe();
58         System.out.println("suspend t1 2 sec");
59         Thread.sleep(2000);
60         System.out.println("resume t1");
61         t1.resumeMe();
62     }
```

在代码第5行，给出一个标记变量suspendme，表示当前线程是否被挂起。同时，增加了suspendMe()和resumeMe()两个方法，分别用于挂起线程和继续执行线程。

在代码第21~28行，线程会先检查自己是否被挂起，如果是，则执行wait()方法进行等待。否则，则进行正常的处理。当线程继续执行时，resumeMe()方法被调用（代码第11~16行），线程t1得到一个继续执行的notify()通知，并且清除了挂起标记，从而得以正常执行。

2.2.6 等待线程结束（**join**）和谦让（**yield**）

在很多情况下，线程之间的协作和人与人之间的协作非常类似。一种非常常见的合作方式就是分工合作。以我们非常熟悉的软件开发为例，在一个项目进行时，总是应该有几位号称是“需求分析师”的同事，先对系统的需求和功能点进行整理和总结，然后，以书面形式给出一份需求说明或者类似的参考文档，然后，软件设计师、研发工程师才会一拥而上，进行软件开发。如果缺少需求分析师的工作输出，那么软件研发的难度可能会比较大。因此，作为一名软件研发人员，总是喜欢等待需求分析师完成他应该完成的任务后，才愿意投身工作。简单地说，就是研发人员需要等待需求分析师完成他的工作，然后，才能进行研发。

将这个关系对应到多线程应用中，很多时候，一个线程的输入可能非常依赖于另外一个或者多个线程的输出，此时，这个线程就需要等待依赖线程执行完毕，才能继续执行。JDK提供了join()操作来实现这个功能，如下所示，显示了2个join()方法：

```
public final void join() throws InterruptedException
public final synchronized void join(long millis) throws Interrupt
```

第一个join()方法表示无限等待，它会一直阻塞当前线程，直到目标线程执行完毕。第二个方法给出了一个最大等待时间，如果超过给定时间目标线程还在执行，当前线程也会因为“等不及了”，而继续往下执行。

英文join的翻译，通常是加入的意思。在这里感觉也非常贴切。因为一个线程要加入另外一个线程，那么最好的方法就是等着它一起走。

这里提供一个简单点的join()实例，供大家参考：

```
public class JoinMain {
    public volatile static int i=0;
    public static class AddThread extends Thread{
        @Override
        public void run() {
            for(i=0;i<100000000;i++);
        }
    }
    public static void main(String[] args) throws InterruptedException {
        AddThread at=new AddThread();
        at.start();
        at.join();
        System.out.println(i);
    }
}
```

主函数中，如果不使用`join()`等待`AddThread`，那么得到的`i`很可能是0或者一个非常小的数字。因为`AddThread`还没开始执行，`i`的值就已经被输出了。但在使用`join()`方法后，表示主线程愿意等待`AddThread`执行完毕，跟着`AddThread`一起往前走，故在`join()`返回时，`AddThread`已经执行完成，故`i`总是10000000。

有关`join()`，我还想再补充一点，`join()`的本质是让调用线程`wait()`在当前线程对象实例上。下面是JDK中`join()`实现的核心代码片段：

```
while (isAlive()) {  
    wait(0);  
}
```

可以看到，它让调用线程在当前线程对象上进行等待。当线程执行完成后，被等待的线程会在退出前调用`notifyAll()`通知所有的等待线程继续执行。因此，值得注意的一点是：不要在应用程序中，在`Thread`对象实例上使用类似`wait()`或者`notify()`等方法，因为这很有可能会影响系统API的工作，或者被系统API所影响。

另外一个比较有趣的方法，是`Thread.yield()`，它的定义如下：

```
public static native void yield();
```

这是一个静态方法，一旦执行，它会使当前线程让出CPU。但要注意，让出CPU并不表示当前线程不执行了。当前线程在让出CPU后，还会进行CPU资源的争夺，但是是否能够再次被分配到，就不一定了。因此，对`Thread.yield()`的调用就好像是在说：我已经完成一些最重要的工作了，我应该是可以休息一下了，可以给其他线程一些工作机会啦！

如果你觉得一个线程不那么重要，或者优先级非常低，而且又害怕它会占用太多的CPU资源，那么可以在适当的时候调用`Thread.yield()`，给予其他重要线程更多的工作机会。

2.3 volatile与Java内存模型（JMM）

之前已经简单介绍了Java内存模型（JMM），Java内存模型都是围绕着原子性、有序性和可见性展开的。大家可以先回顾一下上一章中的相关内容。为了在适当的场合，确保线程间的有序性、可见性和原子性。Java使用了一些特殊的操作或者关键字来申明、告诉虚拟机，在这个地方，要尤其注意，不能随意变动优化目标指令。关键字volatile就是其中之一。

如果你查阅一下英文字典，有关volatile的解释，你会得到最常用的解释是“易变的，不稳定的”。这也正是使用volatile关键字的语义。

当你用volatile去申明一个变量时，就等于告诉了虚拟机，这个变量极有可能会被某些程序或者线程修改。为了确保这个变量被修改后，应用程序范围内的所有线程都能够“看到”这个改动，虚拟机就必须采用一些特殊的手段，保证这个变量的可见性等特点。

比如，根据编译器的优化规则，如果不使用volatile申明变量，那么这个变量被修改后，其他线程可能并不会被通知到，甚至在别的线程中，看到变量的修改顺序都会是反的。但一旦使用volatile，虚拟机就会特别小心地处理这种情况。

大家应该对上一章中介绍原子性时，给出的MultiThreadLong案例还记忆犹新吧！我想，没有人愿意就这么把数据“写坏”。那这种情况，应该怎么处理才能保证每次写进去的数据不坏呢？最简单的一种方法就是

加入volatile申明，告诉编译器，这个long型数据，你要格外小心，因为他会不断地被修改。

下面的代码片段显示了volatile的使用，限于篇幅，这里不再给出完整代码：

```
public class MultiThreadLong {  
    public volatile static long t=0;  
    public static class ChangeT implements Runnable{  
        private long to;  
.....
```

从这个案例中，我们可以看到，volatile对于保证操作的原子性是有非常大的帮助的。但是需要注意的是，volatile并不能代替锁，它也无法保证一些复合操作的原子性。比如下面的例子，通过volatile是无法保证i++的原子性操作的：

```
01 static volatile int i=0;  
02 public static class PlusTask implements Runnable{  
03     @Override  
04     public void run() {  
05         for(int k=0;k<10000;k++)  
06             i++;  
07     }  
08 }  
09  
10 public static void main(String[] args) throws InterruptedException  
11     Thread[] threads=new Thread[10];
```

```
12     for(int i=0;i<10;i++){
13         threads[i]=new Thread(new PlusTask());
14         threads[i].start();
15     }
16     for(int i=0;i<10;i++){
17         threads[i].join();
18     }
19
20     System.out.println(i);
21 }
```

执行上述代码，如果第6行*i++*是原子性的，那么最终的值应该是100000（10个线程各累加10000次）。但实际上，上述代码的输出总是会小于100000。

此外，`volatile`也能保证数据的可见性和有序性。下面再来看一个简单的例子：

```
01 public class NoVisibility {
02     private static boolean ready;
03     private static int number;
04
05     private static class ReaderThread extends Thread {
06         public void run() {
07             while (!ready);
08             System.out.println(number);
09         }
10     }
```

```
11
12     public static void main(String[] args) throws InterruptedException
13         new ReaderThread().start();
14         Thread.sleep(1000);
15         number = 42;
16         ready = true;
17         Thread.sleep(10000);
18     }
19 }
```

上述代码中，ReaderThread线程只有在数据准备好时（ready为true），才会打印number的值。它通过ready变量判断是否应该打印。在主线程中，开启ReaderThread后，就为number和ready赋值，并期望ReaderThread能够看到这些变化并将数据输出。

在虚拟机的Client模式下，由于JIT并没有做足够的优化，在主线程修改ready变量的状态后，ReaderThread可以发现这个改动，并退出程序。但是在Server模式下，由于系统优化的结果，ReaderThread线程无法“看到”主线程中的修改，导致ReaderThread永远无法退出（因为代码第7行判断永远不会成立），这显然不是我们想看到的结果。这个问题就是一个典型的可见性问题。

注意：可以使用Java虚拟机参数-server切换到Server模式。

和原子性问题一样，我们只要简单地使用volatile来申明ready变量，告诉Java虚拟机，这个变量可能会在不同的线程中修改。这样，就可以顺利解决这个问题了。

2.4 分门别类的管理：线程组

在一个系统中，如果线程数量很多，而且功能分配比较明确，就可以将相同功能的线程放置在一个线程组里。打个比方，如果你有一个苹果，你就可以把它拿在手里，但是如果你有十个苹果，你就最好还有一个篮子，否则不方便携带。对于多线程来说，也是这个道理。想要轻松处理几十个甚至上百个线程，最好还是将它们都装进对应的篮子里。

线程组的使用非常简单，如下：

```
01 public class ThreadGroupName implements Runnable {
02     public static void main(String[] args) {
03         ThreadGroup tg = new ThreadGroup("PrintGroup");
04         Thread t1 = new Thread(tg, new ThreadGroupName(), "T1"
05         Thread t2 = new Thread(tg, new ThreadGroupName(), "T2"
06         t1.start();
07         t2.start();
08         System.out.println(tg.activeCount());
09         tg.list();
10     }
11
12     @Override
13     public void run() {
14         String groupAndName=Thread.currentThread().getThreadGr
15             + "-" + Thread.currentThread().getName();
16         while (true) {
```

```
17         System.out.println("I am " + groupAndName);
18         try {
19             Thread.sleep(3000);
20         } catch (InterruptedException e) {
21             e.printStackTrace();
22         }
23     }
24 }
25 }
```

上述代码第3行，建立一个名为“PrintGroup”的线程组，并将T1和T2两个线程加入这个组中。第8、9两行，展示了线程组的两个重要的功能，`activeCount()`可以获得活动线程的总数，但由于线程是动态的，因此这个值只是一个估计值，无法确定精确，`list()`方法可以打印这个线程组中所有的线程信息，对调试有一定帮助。代码中第4、5两行创建了两个线程，使用`Thread`的构造函数，指定线程所属的线程组，将线程和线程组关联起来。

线程组还有一个值得注意的方法`stop()`，它会停止线程组中所有的线程。这看起来是一个很方便的功能，但是它会遇到和`Thread.stop()`相同的问题，因此使用时也需要格外谨慎。

此外，对于编码习惯，我还想再多说几句。强烈建议大家在创建线程和线程组的时候，给它们取一个好听的名字。对于计算机来说，也许名字并不重要，但是在系统出现问题时，你很有可能会导出系统内所有线程，你拿到的如果是一连串的`Thread-0`、`Thread-1`、`Thread-2`，我想你一定会抓狂。但取而代之，你看到的如果是类似`HttpHandler`、`FTPService`这样的名字，会让你心情倍爽。

2.5 驻守后台：守护线程 (Daemon)

守护线程是一种特殊的线程，就和它的名字一样，它是系统的守护者，在后台默默地完成一些系统性的服务，比如垃圾回收线程、JIT线程就可以理解为守护线程。与之相对应的是用户线程，用户线程可以认为是系统的工作线程，它会完成这个程序应该要完成的业务操作。如果用户线程全部结束，这也意味着这个程序实际上无事可做了。守护线程要守护的对象已经不存在了，那么整个应用程序就自然应该结束。因此，当一个Java应用内，只有守护线程时，Java虚拟机就会自然退出。

下面简单地看一下守护线程的使用：

```
01 public class DaemonDemo {
02     public static class DaemonT extends Thread{
03         public void run(){
04             while(true){
05                 System.out.println("I am alive");
06                 try {
07                     Thread.sleep(1000);
08                 } catch (InterruptedException e) {
09                     e.printStackTrace();
10                 }
11             }
12         }
13     }
14 }
```

```
13     }
14     public static void main(String[] args) throws InterruptedException
15         Thread t=new DaemonT();
16         t.setDaemon(true);
17         t.start();
18
19         Thread.sleep(2000);
20     }
21 }
```

上述代码第16行，将线程t设置为守护线程。这里注意，设置守护线程必须在线程start()之前设置，否则你会得到一个类似以下的异常，告诉你守护线程设置失败。但是你的程序和线程依然可以正常执行。只是被当做用户线程而已。因此，如果不小心忽略了下面的异常信息，你就很可能察觉不到这个错误。那你就会诧异为什么程序永远停不下来了呢？

```
Exception in thread "main" java.lang.IllegalThreadStateException
    at java.lang.Thread.setDaemon(Thread.java:1367)
    at geym.conc.ch2.daemon.DaemonDemo.main(DaemonDemo.java:20)
```

在这个例子中，由于t被设置为守护线程，系统中只有主线程main为用户线程，因此在main线程休眠2秒后退出时，整个程序也随之结束。但如果不把线程t设置为守护线程，main线程结束后，t线程还会不停地打印，永远不会结束。

2.6 先干重要的事：线程优先级

Java中的线程可以有自己的优先级。优先级高的线程在竞争资源时会更有优势，更可能抢占资源，当然，这只是一个概率问题。如果运气不好，高优先级线程可能也会抢占失败。由于线程的优先级调度和底层操作系统有密切的关系，在各个平台上表现不一，并且这种优先级产生的后果也可能不容易预测，无法精准控制，比如一个低优先级的线程可能一直抢占不到资源，从而始终无法运行，而产生饥饿（虽然优先级低，但是也不能饿死它呀）。因此，在要求严格的场合，还是需要自己在应用层解决线程调度问题。

在Java中，使用1到10表示线程优先级。一般可以使用内置的三个静态标量表示：

```
public final static int MIN_PRIORITY = 1;
public final static int NORM_PRIORITY = 5;
public final static int MAX_PRIORITY = 10;
```

数字越大则优先级越高，但有效范围在1到10之间。下面的代码展示了优先级的作用。高优先级的线程倾向于更快地完成。

```
01 public class PriorityDemo {
02     public static class HightPriority extends Thread{
03         static int count=0;
04         public void run(){
05             while(true){
```

```
06         synchronized(PriorityDemo.class){
07             count++;
08             if(count>100000000){
09                 System.out.println("HightPriority is c
10                 break;
11             }
12         }
13     }
14 }
15 }
16 public static class LowPriority extends Thread{
17     static int count=0;
18     public void run(){
19         while(true){
20             synchronized(PriorityDemo.class){
21                 count++;
22                 if(count>100000000){
23                     System.out.println("LowPriority is com
24                     break;
25                 }
26             }
27         }
28     }
29 }
30
31 public static void main(String[] args) throws InterruptedE
32     Thread high=new HightPriority();
```

```
33         LowPriority low=new LowPriority();
34         high.setPriority(Thread.MAX_PRIORITY);
35         low.setPriority(Thread.MIN_PRIORITY);
36         low.start();
37         high.start();
38     }
39 }
```

上述代码定义两个线程，分别为HightPriority设置为高优先级，LowPriority为低优先级。让它们完成相同的工作，也就是把count从0加到10000000。完成后，打印信息给一个提示，这样我们就知道谁先完成工作了。这里要注意，在对count累加前，我们使用synchronized产生了一次资源竞争。目的是使得优先级的差异表现得更为明显。

大家可以尝试执行上述代码，可以看到，高优先级的线程在大部分情况下，都会首先完成任务（就这段代码而言，试运行多次，HightPriority总是比LowPriority快，但这不能保证在所有情况下，一定都是这样）。

2.7 线程安全的概念与 synchronized

并行程序开发的一大关注重点就是线程安全。一般来说，程序并行化是为了获得更高的执行效率，但前提是，高效率不能以牺牲正确性为代价。如果程序并行化后，连基本的执行结果的正确性都无法保证，那么并行程序本身也就没有任何意义了。因此，线程安全就是并行程序的根本和根基。大家还记得那个多线程读写long型数据的案例吧！这就是一个典型的反例。但在使用volatile关键字后，这种情况有所改善。但是，volatile并不能真正的保证线程安全。它只能确保一个线程修改了数据后，其他线程能够看到这个改动。但当两个线程同时修改某一个数据时，却依然会产生冲突。

下面的代码演示了一个计数器，两个线程同时对i进行累加操作，各执行10000000次。我们希望的执行结果当然是最终i的值可以达到20000000，但事实并非总是如此。如果你多执行几次下述代码，你会发现，在很多时候，i的最终值会小于20000000。这就是因为两个线程同时对i进行写入时，其中一个线程的结果会覆盖另外一个（虽然这个时候i被声明为volatile变量）。

```
01 public class AccountingVol implements Runnable{
02     static AccountingVol instance=new AccountingVol();
03     static volatile int i=0;
04     public static void increase(){
05         i++;
```

```

06     }
07     @Override
08     public void run() {
09         for(int j=0;j<10000000;j++){
10             increase();
11         }
12     }
13     public static void main(String[] args) throws InterruptedException
14         Thread t1=new Thread(instance);
15         Thread t2=new Thread(instance);
16         t1.start();t2.start();
17         t1.join();t2.join();
18         System.out.println(i);
19     }
20 }

```

图2.8展示了这种可能的冲突，如果在代码中发生了类似的情况，这就是多线程不安全的恶果。线程1和线程2同时读取*i*为0，并各自计算得到*i*=1，并先后写入这个结果，因此，虽然*i*++被执行了2次，但是实际*i*的值只增加了1。

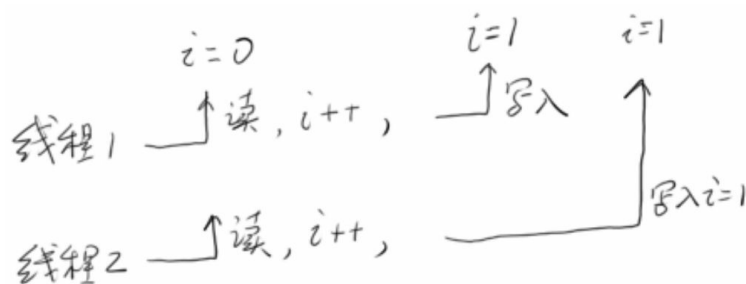


图2.8 多线程的写入冲突

要从根本上解决这个问题，我们就必须保证多个线程在对*i*进行操作时完全同步。也就是说，当线程A在写入时，线程B不仅不能写，同时也不能读。因为在线程A写完之前，线程B读取的一定是一个过期数据。Java中，提供了一个重要的关键字synchronized来实现这个功能。

关键字synchronized的作用是实现线程间的同步。它的工作是对同步的代码加锁，使得每一次，只能有一个线程进入同步块，从而保证线程间的安全性（也就是说在上述代码的第5行，每次应该只有一个线程可以执行）。

关键字synchronized可以有多种用法。这里做一个简单的整理。

- 指定加锁对象：对给定对象加锁，进入同步代码前要获得给定对象的锁。
- 直接作用于实例方法：相当于对当前实例加锁，进入同步代码前要获得当前实例的锁。
- 直接作用于静态方法：相当于对当前类加锁，进入同步代码前要获得当前类的锁。

下述代码，将synchronized作用于一个给定对象instance，因此，每次当线程进入被synchronized包裹的代码段，就都会要求请求instance实例的锁。如果当前有其他线程正持有这把锁，那么新到的线程就必须等待。这样，就保证了每次只能有一个线程执行*i*++操作。

```
public class AccountingSync implements Runnable{
    static AccountingSync instance=new AccountingSync();
    static int i=0;
    @Override
```

```

public void run() {
    for(int j=0;j<10000000;j++){
        synchronized(instance){
            i++;
        }
    }
}

```

//main函数参见本节第一段代码

当然，上述代码也可以写成如下形式，两者是等价的：

```

01 public class AccountingSync2 implements Runnable{
02     static AccountingSync2 instance=new AccountingSync2();
03     static int i=0;
04     public synchronized void increase(){
05         i++;
06     }
07     @Override
08     public void run() {
09         for(int j=0;j<10000000;j++){
10             increase();
11         }
12     }
13     public static void main(String[] args) throws InterruptedException
14         Thread t1=new Thread(instance);
15         Thread t2=new Thread(instance);
16         t1.start();t2.start();

```

```
17         t1.join();t2.join();
18         System.out.println(i);
19     }
20 }
```

上述代码中，`synchronized`关键字作用于一个实例方法。这就是说在进入`increase()`方法前，线程必须获得当前对象实例的锁。在本例中就是`instance`对象。在这里，我不厌其烦地再次给出`main`函数的实现，是希望强调第14、15行代码，也就是`Thread`的创建方式。这里使用`Runnable`接口创建两个线程，并且这两个线程都指向同一个`Runnable`接口实例（`instance`对象），这样才能保证两个线程在工作时，能够关注到同一个对象锁上去，从而保证线程安全。

一种错误的同步方式如下：

```
01 public class AccountingSyncBad implements Runnable{
02     static int i=0;
03     public synchronized void increase(){
04         i++;
05     }
06     @Override
07     public void run() {
08         for(int j=0;j<10000000;j++){
09             increase();
10         }
11     }
12     public static void main(String[] args) throws InterruptedException
13         Thread t1=new Thread(new AccountingSyncBad());
```



```
14      Thread t2=new Thread(new AccountingSyncBad());
15      t1.start();t2.start();
16      t1.join();t2.join();
17      System.out.println(i);
18  }
19 }
```

上述代码就犯了一个严重的错误。虽然在第3行的`increase()`方法中，申明这是一个同步方法。但很不幸的是，执行这段代码的两个线程都指向了不同的`Runnable`实例。由第13、14行可以看到，这两个线程的`Runnable`实例并不是同一个对象。因此，线程t1会在进入同步方法前加锁自己的`Runnable`实例，而线程t2也关注于自己的对象锁。换言之，这两个线程使用的是两把不同的锁。因此，线程安全是无法保证的。

但我们只要简单地修改上述代码，就能使其正确执行。那就是使用`synchronized`的第三种用法，将其作用于静态方法。将`increase()`方法修改如下：

```
public static synchronized void increase(){
    i++;
}
```

这样，即使两个线程指向不同的`Runnable`对象，但由于方法块需要请求的是当前类的锁，而非当前实例，因此，线程间还是可以正确同步。

除了用于线程同步、确保线程安全外，`synchronized`还可以保证线程间的可见性和有序性。从可见性的角度上讲，`synchronized`可以完全

替代volatile的功能，只是使用上没有那么方便。就有序性而言，由于synchronized限制每次只有一个线程可以访问同步块，因此，无论同步块内的代码如何被乱序执行，只要保证串行语义一致，那么执行结果总是一样的。而其他访问线程，又必须在获得锁后方能进入代码块读取数据，因此，它们看到的最终结果并不取决于代码的执行过程，从而有序性问题自然得到了解决（换言之，被synchronized限制的多个线程是串行执行的）。

2.8 程序中的幽灵：隐蔽的错误

作为一名软件开发人员，修复程序BUG应该说是基本的日常工作之一。作为Java程序员，也许你经常会被抛出的一大堆的异常堆栈所困扰，因为这可能预示着你又有工作可做了。但我这里想说的是，如果程序出错，你看到了异常堆栈，那你应该感到额外的高兴，因为这也意味着你极有可能可以在两分钟内修复这个问题（当然，并不是所有的异常都是错误）。最可怕的情况是：系统没有任何异常表现，没有日志，也没有堆栈，但是却给出了一个错误的执行结果，这种情况下，才真会让你抓狂。

2.8.1 无提示的错误案例

我在这里想给出一个系统运行错误，却没有任何提示的案例。让大家体会一下这种情况的可怕之处。我相信，在任何一个业务系统中，求平均值，应该是一种极其常见的操作。这里就以求两个整数的平均值为例。请看下面代码：

```
int v1=1073741827;
int v2=1431655768;
System.out.println("v1="+v1);
System.out.println("v2="+v2);
int ave=(v1+v2)/2;
System.out.println("ave="+ave);
```

上述代码中，加粗部分试图计算v1和v2的均值。乍看之下，没有什么问题。目测v1和v2的当前值，估计两者的平均值大约在12亿左右。但如果你执行代码，却会得到以下输出：

```
v1=1073741827
v2=1431655768
ave=-894784850
```

乍看之下，你一定会觉得非常吃惊，为什么均值竟然反而是一个负数。但只要你有一点研发精神，就会马上有所觉悟。这是一个典型的溢出问题！显然，v1+v2的结果就已经导致了int的溢出。

把这个问题单独拿出来研究，也许你不会有特别的感触，但是，一旦这个问题发生在一个复杂系统的内部。由于复杂的业务逻辑，很可能掩盖这个看起来微不足道的问题，再加上程序自始至终没有任何日志或异常，再加上你运气不是太好的话，这类问题不让你耗上几个通宵，恐怕也是难有眉目。

所以，我们自然会恐惧这些问题，我们也希望在程序异常时，能够得到一个异常或者相关的日志。但是，非常不幸的是，错误地使用并行，会非常容易产生这类问题。它们难觅踪影，就如同幽灵一般。

2.8.2 并发下的ArrayList

我们都知道，ArrayList是一个线程不安全的容器。如果在多线程中使用ArrayList，可能会导致程序出错。那究竟可能引起哪些问题呢？试看下面的代码：

```

public class ArrayListMultiThread {
    static ArrayList<Integer> al = new ArrayList<Integer>(10);
    public static class AddThread implements Runnable {
        @Override
        public void run() {
            for (int i = 0; i < 10000000; i++) {
                al.add(i);
            }
        }
    }

    public static void main(String[] args) throws InterruptedException {
        Thread t1=new Thread(new AddThread());
        Thread t2=new Thread(new AddThread());
        t1.start();
        t2.start();
        t1.join();t2.join();
        System.out.println(al.size());
    }
}

```

上述代码中，t1和t2两个线程同时向一个ArrayList容器中添加元素。他们各添加10000000个元素，因此我们期望最后可以有20000000个元素在ArrayList中。但如果你执行这段代码，你可能会得到三种结果。

第一，程序正常结束，ArrayList的最终大小确实20000000。这说明即使并程序有问题，也未必会每次都表现出来。

第二，程序抛出异常：

```
Exception in thread "Thread-0" java.lang.ArrayIndexOutOfBoundsException  
    at java.util.ArrayList.add(ArrayList.java:441)  
    at geym.conc.ch2.notsafe.ArrayListMultiThread$AddThread.run  
(ArrayListMultiThread.java:12)  
    at java.lang.Thread.run(Thread.java:724)  
1000015
```

这是因为ArrayList在扩容过程中，内部一致性被破坏，但由于没有锁的保护，另外一个线程访问到了不一致的内部状态，导致出现越界问题。

第三，出现了一个非常隐蔽的错误，比如打印如下值作为ArrayList的大小：

```
1793758
```

显然，这是由于多线程访问冲突，使得保存容器大小的变量被多线程不正常的访问，同时两个线程也同时对ArrayList中的同一个位置进行赋值导致的。如果出现这种问题，那么很不幸，你就得到了一个没有错误提示的错误。并且，他们未必是可以复现的。

注意：改进的方法很简单，使用线程安全的Vector代替ArrayList即可。

2.8.3 并发下诡异的HashMap

HashMap同样不是线程安全的。当你使用多线程访问HashMap时，也可能会遇到意想不到的错误。不过和ArrayList不同，HashMap的问题似乎更加诡异。

```
public class HashMapMultiThread {

    static Map<String,String> map = new HashMap<String,String>();

    public static class AddThread implements Runnable {
        int start=0;
        public AddThread(int start){
            this.start=start;
        }
        @Override
        public void run() {
            for (int i = start; i < 100000; i+=2) {
                map.put(Integer.toString(i), Integer.toBinaryString(i));
            }
        }
    }

    public static void main(String[] args) throws InterruptedException {
        Thread t1=new Thread(new HashMapMultiThread.AddThread(0))
        Thread t2=new Thread(new HashMapMultiThread.AddThread(1))
        t1.start();
        t2.start();
        t1.join();t2.join();
    }
}
```

```
        System.out.println(map.size());  
    }  
}
```

上述代码使用t1和t2两个线程同时对HashMap进行put()操作。如果一切正常，我们期望得到的map.size()就是100000。但实际上，你可能会得到以下三种情况（注意，这里使用JDK 7进行试验）：

第一，程序正常结束，并且结果也是符合预期的。HashMap的大小为100000。

第二，程序正常结束，但结果不符合预期，而是一个小于100000的数字，比如98868。

第三，程序永远无法结束。

对于前两种可能，和ArrayList的情况非常类似，因此，也不必过多解释。而对于第三种情况，如果是第一次看到，我想大家一定会觉得特别惊讶，因为看似非常正常的程序，怎么可能就结束不了呢？

注意：请读者谨慎尝试以上代码，由于这段代码很可能占用两个CPU核，并使它们的CPU占有率达到100%。如果CPU性能较弱，可能导致死机。请先保存资料，再进行尝试。

打开任务管理器，你们会发现，这段代码占用了极高的CPU，最有可能的表示是占用了两个CPU核，并使得这两个核的CPU使用率达到100%。这非常类似死循环的情况。

使用jstack工具显示程序的线程信息，如下所示。其中jps可以显示当前系统中所有的Java进程。而jstack可以打印给定Java进程的内部线程

及其堆栈。

```
C:\Users\geym >jps
14240 HashMapMultiThread
1192 Jps
C:\Users\geym >jstack 14240
```

我们会很容易找到我们的t1、t2和main线程：

```
"Thread-1" prio=6 tid=0x00bb2800 nid=0x16e0 runnable [0x04baf000]
  java.lang.Thread.State: RUNNABLE
    at java.util.HashMap.put(HashMap.java:498)
    at geym.conc.ch2.notsafe.HashMapMultiThread$AddThread.run
(HashMapMultiThread.java:26)
    at java.lang.Thread.run(Thread.java:724)

"Thread-0" prio=6 tid=0x00bb0000 nid=0x1668 runnable [0x04d7f000]
  java.lang.Thread.State: RUNNABLE
    at java.util.HashMap.put(HashMap.java:498)
    at geym.conc.ch2.notsafe.HashMapMultiThread$AddThread.run
(HashMapMultiThread.java:26)
    at java.lang.Thread.run(Thread.java:724)

"main" prio=6 tid=0x00c0cc00 nid=0x16ec in Object.wait() [0x0102f
  java.lang.Thread.State: WAITING (on object monitor)
    at java.lang.Object.wait(Native Method)
    - waiting on <0x24930280> (a java.lang.Thread)
    at java.lang.Thread.join(Thread.java:1260)
    - locked <0x24930280> (a java.lang.Thread)
```

```
at java.lang.Thread.join(Thread.java:1334)
at geym.conc.ch2.otsafe.HashMapMultiThread.main(HashMapM
```

可以看到，主线程main正处于等待状态，并且这个等待是由于join()方法引起的，符合我们的预期。而t1和t2两个线程都处于Runnable状态，并且当前执行语句为HashMap.put()方法。查看put()方法的第498行代码，如下所示：

```
498 for (Entry<K,V> e = table[i]; e != null; e = e.next) {
499     Object k;
500     if (e.hash == hash && ((k = e.key) == key || key.equals(k
501         V oldValue = e.value;
502         e.value = value;
503         e.recordAccess(this);
504         return oldValue;
505     }
506 }
```

可以看到，当前这两个线程正在遍历HashMap的内部数据。当前所处循环乍看之下是一个迭代遍历，就如同遍历一个链表一样。但在此时此刻，由于多线程的冲突，这个链表的结构已经遭到了破坏，链表成环了！当链表成环时，上述的迭代就等同于一个死循环，如图2.9所示，展示了最简单的一种环状结构，Key1和Key2互为对方的next元素。此时，通过next引用遍历，将形成死循环。

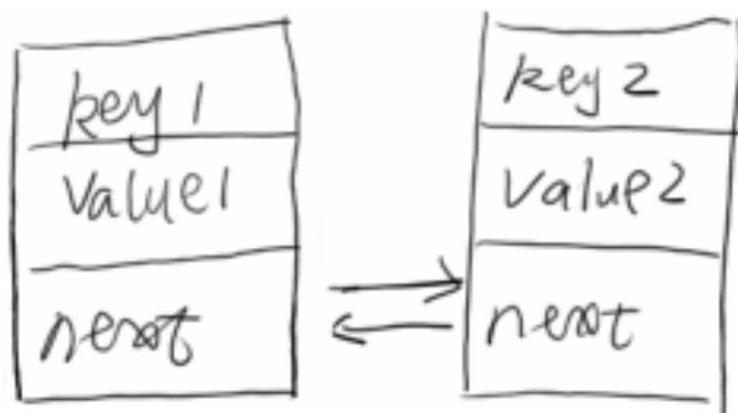


图2-9 成环的链表

这个死循环的问题，如果一旦发生，着实可以让你郁闷一把。本章的参考资料中也给出了一个真实的案例。但这个死循环的问题在JDK 8中已经不存在了。由于JDK 8对HashMap的内部实现了做了大规模的调整，因此规避了这个问题。但即使这样，贸然在多线程环境下使用HashMap依然会导致内部数据不一致。最简单的解决方案就是使用ConcurrentHashMap代替HashMap。

2.8.4 初学者常见问题：错误的加锁

在进行多线程同步时，加锁是保证线程安全的重要手段之一。但加锁也必须是合理的，在“线程安全的概念与synchronized”一节中，我已经给出了一个常见的错误加锁的案例。也就是锁的不正确使用。在本节中，我将介绍一个更加隐晦的错误。

现在，假设我们需要一个计数器，这个计数器会被多个线程同时访问。为了确保数据正确性，我们自然会需要对计数器加锁，因此，就有了以下代码：

```
01 public class BadLockOnInteger implements Runnable{
```

```

02     public static Integer i=0;
03     static BadLockOnInteger instance=new BadLockOnInteger();
04     @Override
05     public void run() {
06         for(int j=0;j<100000000;j++){
07             synchronized(i){
08                 i++;
09             }
10         }
11     }
12
13     public static void main(String[] args) throws InterruptedException
14     {
15         Thread t1=new Thread(instance);
16         Thread t2=new Thread(instance);
17         t1.start();t2.start();
18         t1.join();t2.join();
19         System.out.println(i);
20     }

```

上述代码的第7~9行，为了保证计数器*i*的正确性，每次对*i*自增前，都先获得*i*的锁，以此保证*i*是线程安全的。从逻辑上看，这似乎并没有什么不对，所以，我们就满怀信心地尝试运行我们的代码。如果一切正常，这段代码应该返回200000000（每个线程各累加100000000次）。

但结果却让我们惊呆了，我得到了一个比200000000小很多的数字，比如15992526。这说明什么问题呢？一定是这段程序并没有真正做到线

程安全！但把锁加在变量*i*上又有什么问题呢？似乎加锁的逻辑也是无懈可击的。

要解释这个问题，得从Integer说起。在Java中，Integer属于不变对象。也就是对象一旦被创建，就不可能被修改。也就是说，如果你有一个Integer代表1，那么它就永远表示1，你不可能修改Integer的值，使它为2。那如果你需要2怎么办呢？也很简单，新建一个Integer，并让它表示2即可。

如果我们使用javap反编译这段代码的run()方法，我们可以看到：

```
0:   iconst_0
1:   istore_1
2:   goto    36
5:   getstatic      #20; //Field i:Ljava/lang/Integer;
8:   dup
9:   astore_2
10:  monitorenter
11:  getstatic      #20; //Field i:Ljava/lang/Integer;
14:  invokevirtual  #32; //Method java/lang/Integer.intValue:()I
17:  iconst_1
18:  iadd
19:  invokestatic   #14; //Method java/lang/Integer.valueOf:(I)L
22:  putstatic      #20; //Field i:Ljava/lang/Integer;
25:  aload_2
26:  monitorexit
```

在第19~22行（对字节码来说，这是偏移量，这里简称为行），实

际上使用了Integer.valueOf()方法新建了一个新的Integer对象，并将它赋值给变量i。也就是说，i++在真实执行时变成了：

```
i=Integer.valueOf(i.intValue()+1);
```

进一步查看Integer.valueOf()，我们可以看到：

```
public static Integer valueOf(int i) {  
    assert IntegerCache.high >= 127;  
    if (i >= IntegerCache.low && i <= IntegerCache.high)  
        return IntegerCache.cache[i + (-IntegerCache.low)];  
    return new Integer(i);  
}
```

Integer.valueOf()实际上是一个工厂方法，它会倾向于返回一个代表指定数值的Integer实例。因此，i++的本质是，创建一个新的Integer对象，并将它的引用赋值给i。

如此一来，我们就可以明白问题所在了，由于在多个线程间，并不一定能够看到同一个i对象（因为i对象一直在变），因此，两个线程每次加锁可能都加在了不同的对象实例上，从而导致对临界区代码控制出现问题。

修正这个问题也很容易，只要将

```
synchronized(i){
```

改为：

```
synchronized(instance){
```

即可。

2.9 参考文献

- 这篇文章生动形象地描述了线程和进程
 - http://www.qnx.com/developers/docs/6.4.1/neutrino/getting_started/
- 有关线程的状态机
 - <http://www.cnblogs.com/skywang12345/p/3479024.html>
- 对线程中断给出极其详细的描述
 - <http://ibruce.info/2013/12/19/how-to-stop-a-java-thread/>
- 对Java虚拟机的Server和Client模式进行了说明
 - <http://www.uucode.net/201406/jvm-server-client-mode>
- 线程组的概念与使用
 - <http://ifeve.com/thread-management-11/>
- 有关守护线程详尽描述
 - <http://blog.csdn.net/lcore/article/details/12280027>
- HashMap在多线程卡死的细节分析
 - <http://coolshell.cn/articles/9606.html>

- WeakHashMap多线程卡死的实际案例
 - <http://www.uucode.net/201412/weakhashmap-endless-loop>
- HashMap的实现原理
 - <http://www.uucode.net/201503/hashmap-hash-col>

第3章 JDK并发包

为了更好地支持并发程序，JDK内部提供了大量实用的API和框架。在本章中，将主要介绍这些JDK内部的功能，其主要分为三大部分：

首先，将介绍有关同步控制的工具，之前介绍的synchronized关键字就是一种同步控制手段，在这里，我们将看到更加丰富多彩的多线程控制方法。

其次，将详细介绍JDK中对线程池的支持，使用线程池，将能很大程度上提高线程调度的性能。

第三，我将向大家介绍JDK的一些并发容器，这些容器专为并行访问所设计，绝对是高效、安全、稳定的实用工具。

3.1 多线程的团队协作：同步控制

同步控制是并发程序必不可少的重要手段。之前介绍的 `synchronized` 关键字就是一种最简单的控制方法。它决定了一个线程是否可以访问临界区资源。同时，`Object.wait()` 和 `Object.notify()` 方法起到了线程等待和通知的作用。这些工具对于实现复杂的多线程协作起到了重要的作用。在本节中，我们首先将介绍 `synchronized`、`Object.wait()` 和 `Object.notify()` 方法的替代品（或者说是增强版）——重入锁。

3.1.1 `synchronized` 的功能扩展：重入锁

重入锁可以完全替代 `synchronized` 关键字。在 JDK 5.0 的早期版本中，重入锁的性能远远好于 `synchronized`，但从 JDK 6.0 开始，JDK 在 `synchronized` 上做了大量的优化，使得两者的性能差距并不大。

重入锁使用 `java.util.concurrent.locks.ReentrantLock` 类来实现。下面是一段最简单的重入锁使用案例：

```
01 public class ReenterLock implements Runnable{
02     public static ReentrantLock lock=new ReentrantLock();
03     public static int i=0;
04     @Override
05     public void run() {
06         for(int j=0;j<10000000;j++){
```

```

07         lock.lock();
08         try{
09             i++;
10         }finally{
11             lock.unlock();
12         }
13     }
14 }
15 public static void main(String[] args) throws InterruptedException
16     {
17         ReentrantLock tl=new ReentrantLock();
18         Thread t1=new Thread(tl);
19         Thread t2=new Thread(tl);
20         t1.start();t2.start();
21         t1.join();t2.join();
22         System.out.println(i);
23     }

```

上述代码第7~12行，使用重入锁保护临界区资源*i*，确保多线程对*i*操作的安全性。从这段代码可以看到，与synchronized相比，重入锁有着显示的操作过程。开发人员必须手动指定何时加锁，何时释放锁。也正因为这样，重入锁对逻辑控制的灵活性要远远好于synchronized。但值得注意的是，在退出临界区时，必须记得释放锁（代码第11行），否则，其他线程就没有机会再访问临界区了。

有些同学可能会对重入锁的名字感到奇怪。锁就叫锁呗，为什么要加上“重入”两个字呢？从类的命名上看，Re-Entrant-Lock翻译成重入锁

也是非常贴切的。之所以这么叫，那是因为这种锁是可以反复进入的。当然，这里的反复仅仅局限于一个线程。上述代码的第7~12行，可以写成下面的形式：

```
lock.lock();
lock.lock();
try{
    i++;
}finally{
    lock.unlock();
    lock.unlock();
}
```

在这种情况下，一个线程连续两次获得同一把锁。这是允许的！如果不允许这么操作，那么同一个线程在第2次获得锁时，将会和自己产生死锁。程序就会“卡死”在第2次申请锁的过程中。但需要注意的是，如果同一个线程多次获得锁，那么在释放锁的时候，也必须释放相同次数。如果释放锁的次数多，那么会得到一个 `java.lang.IllegalMonitorStateException` 异常，反之，如果释放锁的次数少了，那么相当于线程还持有这个锁，因此，其他线程也无法进入临界区。

除了使用上的灵活性外，重入锁还提供了一些高级功能。比如，重入锁可以提供中断处理的能力。

- 中断响应

对于 `synchronized` 来说，如果一个线程在等待锁，那么结果只有两

种情况，要么它获得这把锁继续执行，要么它就保持等待。而使用重入锁，则提供另外一种可能，那就是线程可以被中断。也就是在等待锁的过程中，程序可以根据需要取消对锁的请求。有些时候，这么做是非常有必要的。比如，如果你和朋友约好一起去打球，如果你等了半小时，朋友还没有到，突然接到一个电话，说由于突发情况，不能如约了。那么你一定就扫兴地打道回府了。中断正式提供了一套类似的机制。如果一个线程正在等待锁，那么它依然可以收到一个通知，被告知无须再等待，可以停止工作了。这种情况对于处理死锁是有一定帮助的。

下面的代码产生了一个死锁，但得益于锁中断，我们可以很轻易地解决这个死锁。

```
01 public class IntLock implements Runnable {
02     public static ReentrantLock lock1 = new ReentrantLock();
03     public static ReentrantLock lock2 = new ReentrantLock();
04     int lock;
05     /**
06      * 控制加锁顺序，方便构造死锁
07      * @param lock
08      */
09     public IntLock(int lock) {
10         this.lock = lock;
11     }
12
13     @Override
14     public void run() {
15         try {
```

```
16         if (lock == 1) {
17             lock1.lockInterruptibly();
18             try{
19                 Thread.sleep(500);
20             }catch(InterruptedException e){}
21             lock2.lockInterruptibly();
22         } else {
23             lock2.lockInterruptibly();
24             try{
25                 Thread.sleep(500);
26             }catch(InterruptedException e){}
27             lock1.lockInterruptibly();
28         }
29
30     } catch (InterruptedException e) {
31         e.printStackTrace();
32     } finally {
33         if (lock1.isHeldByCurrentThread())
34             lock1.unlock();
35         if (lock2.isHeldByCurrentThread())
36             lock2.unlock();
37         System.out.println(Thread.currentThread().getId()+
38     }
39 }
40
41 public static void main(String[] args) throws InterruptedE
42     IntLock r1 = new IntLock(1);
```

```
43      IntLock r2 = new IntLock(2);
44      Thread t1 = new Thread(r1);
45      Thread t2 = new Thread(r2);
46      t1.start();t2.start();
47      Thread.sleep(1000);
48      //中断其中一个线程
49      t2.interrupt();
50  }
51 }
```

线程t1和t2启动后，t1先占用lock1，再占用lock2；t2先占用lock2，再请求lock1。因此，很容易形成t1和t2之间的相互等待。在这里，对锁的请求，统一使用lockInterruptibly()方法。这是一个可以对中断进行响应的锁申请动作，即在等待锁的过程中，可以响应中断。

在代码第47行，主线程main处于休眠，此时，这两个线程处于死锁的状态，在代码第49行，由于t2线程被中断，故t2会放弃对lock1的申请，同时释放已获得lock2。这个操作导致t1线程可以顺利得到lock2而继续执行下去。

执行上述代码，将输出：

```
java.lang.InterruptedException
    at java.util.concurrent.locks.AbstractQueuedSynchronizer.
doAcquireInterruptibly(AbstractQueuedSynchronizer.java:898)
    at java.util.concurrent.locks.AbstractQueuedSynchronizer.
acquireInterruptibly(AbstractQueuedSynchronizer.java:1222)
    at java.util.concurrent.locks.ReentrantLock.lockInterruptibly
```



```
(ReentrantLock.java:335)
    at geym.conc.ch3.syncctrl.IntLock.run(IntLock.java:31)
    at java.lang.Thread.run(Thread.java:745)
9:线程退出
8:线程退出
```

可以看到，中断后，两个线程双双退出。但真正完成工作的只有t1。而t2线程则放弃其任务直接退出，释放资源。

- 锁申请等待限时

除了等待外部通知之外，要避免死锁还有另外一种方法，那就是限时等待。依然以约朋友打球为例，如果朋友迟迟不来，又无法联系到他。那么，在等待1~2个小时后，我想大部分人都会扫兴离去。对线程来说也是这样。通常，我们无法判断为什么一个线程迟迟拿不到锁。也许是因为死锁了，也许是因为产生了饥饿。但如果给定一个等待时间，让线程自动放弃，那么对系统来说是有意义的。我们可以使用tryLock()方法进行一次限时的等待。

下面这段代码展示了限时等待锁的使用。

```
01 public class TimeLock implements Runnable{
02     public static ReentrantLock lock=new ReentrantLock();
03     @Override
04     public void run() {
05         try {
06             if(lock.tryLock(5, TimeUnit.SECONDS)){
07                 Thread.sleep(6000);
```

```

08         }else{
09             System.out.println("get lock failed");
10         }
11     } catch (InterruptedException e) {
12         e.printStackTrace();
13     }finally{if(lock.isHeldByCurrentThread()) lock.unlock(
14     }
15     public static void main(String[] args) {
16         TimeLock tl=new TimeLock();
17         Thread t1=new Thread(tl);
18         Thread t2=new Thread(tl);
19         t1.start();
20         t2.start();
21     }
22 }

```

在这里，tryLock()方法接收两个参数，一个表示等待时长，另外一个表示计时单位。这里的单位设置为秒，时长为5，表示线程在这个锁请求中，最多等待5秒。如果超过5秒还没有得到锁，就会返回false。如果成功获得锁，则返回true。

在本例中，由于占用锁的线程会持有锁长达6秒，故另一个线程无法在5秒的等待时间内获得锁，因此，请求锁会失败。

ReentrantLock.tryLock()方法也可以不带参数直接运行。在这种情况下，当前线程会尝试获得锁，如果锁并未被其他线程占用，则申请锁会成功，并立即返回true。如果锁被其他线程占用，则当前线程不会进行等待，而是立即返回false。这种模式不会引起线程等待，因此也不会产

生死锁。下面演示了这种使用方式：

```
01 public class TryLock implements Runnable {
02     public static ReentrantLock lock1 = new ReentrantLock();
03     public static ReentrantLock lock2 = new ReentrantLock();
04     int lock;
05
06     public TryLock(int lock) {
07         this.lock = lock;
08     }
09
10     @Override
11     public void run() {
12         if (lock == 1) {
13             while (true) {
14                 if (lock1.tryLock()) {
15                     try {
16                         try {
17                             Thread.sleep(500);
18                         } catch (InterruptedException e) {
19                         }
20                         if (lock2.tryLock()) {
21                             try {
22                                 System.out.println(Thread.currentThread().getId() + ":My Job done");
23                                 return;
24                             } finally {
25                                 lock2.unlock();
26                             }
27                         }
28                     } finally {
29                         lock1.unlock();
30                     }
31                 }
32             }
33         }
34     }
35 }
```

```
26             lock2.unlock();
27         }
28     }
29     } finally {
30         lock1.unlock();
31     }
32 }
33 }
34 } else {
35     while (true) {
36         if (lock2.tryLock()) {
37             try {
38                 try {
39                     Thread.sleep(500);
40                 } catch (InterruptedException e) {
41                 }
42                 if (lock1.tryLock()) {
43                     try {
44                         System.out.println(Thread.currentThread().getId() + ":My Job done");
45                     } finally {
46                         lock1.unlock();
47                     }
48                 }
49             }
50         }
51     } finally {
52         lock2.unlock();
53     }
54 }
```

```

53         }
54     }
55 }
56 }
57 }
58
59 public static void main(String[] args) throws InterruptedException
60     TryLock r1 = new TryLock(1);
61     TryLock r2 = new TryLock(2);
62     Thread t1 = new Thread(r1);
63     Thread t2 = new Thread(r2);
64     t1.start();
65     t2.start();
66 }
67 }

```

上述代码中，采用了非常容易死锁的加锁顺序。也就是先让t1获得lock1，再让t2获得lock2，接着做反向请求，让t1申请lock2，t2申请lock1。在一般情况下，这会导致t1和t2相互等待，从而引起死锁。

但是使用tryLock()后，这种情况就大大改善了。由于线程不会傻傻地等待，而是不停地尝试，因此，只要执行足够长的时间，线程总是会得到所有需要的资源，从而正常执行（这里以线程同时获得lock1和lock2两把锁，作为其可以正常执行的条件）。在同时获得lock1和lock2后，线程就打印出标志着任务完成的信息“My Job done”。

执行上述代码，等待一会儿（由于线程中包含休眠500毫秒的代码）。最终你还是可以欣喜地看到程序执行完毕，并产生如下输出，表

示两个线程双双正常执行。

```
9:My Job done
```

```
8:My Job done
```

- 公平锁

在大多数情况下，锁的申请都是不公平的。也就是说，线程1首先请求了锁A，接着线程2也请求了锁A。那么当锁A可用时，是线程1可以获得锁还是线程2可以获得锁呢？这是不一定的。系统只是会从这个锁的等待队列中随机挑选一个。因此不能保证其公平性。这就好比买票不排队，大家都乱哄哄得围在售票窗口前，售票员忙得焦头烂额，也顾不及谁先谁后，随便找个人出票就完事了。而公平的锁，则不是这样，它会按照时间的先后顺序，保证先到者先得，后到者后得。公平锁的一大特点是：它不会产生饥饿现象。只要你排队，最终还是可以等到资源的。如果我们使用synchronized关键字进行锁控制，那么产生的锁就是不公平的。而重入锁允许我们对其公平性进行设置。它有一个如下的构造函数：

```
public ReentrantLock(boolean fair)
```

当参数fair为true时，表示锁是公平的。公平锁看起来很优美，但是要实现公平锁必然要求系统维护一个有序队列，因此公平锁的实现成本比较高，性能相对也非常低下，因此，默认情况下，锁是不公平的。如果没有特别的需求，也不需要使⽤公平锁。公平锁和不公平锁在线程调度表现上也是非常不一样的。下面的代码可以很好地突出公平锁的特点：

```
01 public class FairLock implements Runnable {
```

```

02     public static ReentrantLock fairLock = new ReentrantLock(t
03
04     @Override
05     public void run() {
06         while(true){
07             try{
08                 fairLock.lock();
09                 System.out.println(Thread.currentThread().getName(
10             }finally{
11                 fairLock.unlock();
12             }
13         }
14     }
15
16     public static void main(String[] args) throws InterruptedE
17         FairLock r1 = new FairLock();
18         Thread t1=new Thread(r1,"Thread_t1");
19         Thread t2=new Thread(r1,"Thread_t2");
20         t1.start();t2.start();
21     }
22 }

```

上述代码第2行，指定锁是公平的。接着，由两个线程t1和t2分别请求这把锁，并且在得到锁后，进行一个控制台的输出，表示自己得到了锁。在公平锁的情况下，得到输出通常如下所示：

```
Thread_t1 获得锁
```

```
Thread_t2 获得锁
Thread_t1 获得锁
Thread_t2 获得锁
Thread_t1 获得锁
Thread_t2 获得锁
Thread_t1 获得锁
Thread_t2 获得锁
Thread_t1 获得锁
```

由于代码会产生大量输出，这里只截取部分进行说明。在这个输出中，很明显可以看到，两个线程基本上是交替获得锁的，几乎不会发生同一个线程连续多次获得锁的可能，从而公平性也得到了保证。如果不使用公平锁，那么情况会完全不一样，下面是使用非公平锁时的部分输出：

前面还有一段t1连续获得锁的输出

```
Thread_t1 获得锁
Thread_t1 获得锁
Thread_t1 获得锁
Thread_t1 获得锁
Thread_t2 获得锁
Thread_t2 获得锁
Thread_t2 获得锁
Thread_t2 获得锁
Thread_t2 获得锁
```

后面还有一段t2连续获得锁的输出

可以看到，根据系统的调度，一个线程会倾向于再次获取已经持有

的锁，这种分配方式是高效的，但是无公平性可言。

对上面ReentrantLock的几个重要方法整理如下。

- `lock()`: 获得锁，如果锁已经被占用，则等待。
- `lockInterruptibly()`: 获得锁，但优先响应中断。
- `tryLock()`: 尝试获得锁，如果成功，返回`true`，失败返回`false`。
该方法不等待，立即返回。
- `tryLock(long time, TimeUnit unit)`: 在给定时间内尝试获得锁。
- `unlock()`: 释放锁。

就重入锁的实现来看，它主要集中在Java层面。在重入锁的实现中，主要包含三个要素：

第一，是原子状态。原子状态使用CAS操作（在第4章进行详细讨论）来存储当前锁的状态，判断锁是否已经被别的线程持有。

第二，是等待队列。所有没有请求到锁的线程，会进入等待队列进行等待。待有线程释放锁后，系统就能从等待队列中唤醒一个线程，继续工作。

第三，是阻塞原语`park()`和`unpark()`，用来挂起和恢复线程。没有得到锁的线程将会被挂起。有关`park()`和`unpark()`的详细介绍，可以参考3.1.7线程阻塞工具类：LockSupport。

3.1.2 重入锁的好搭档：Condition条件

如果大家理解了Object.wait()和Object.notify()方法的话，那么就能很容易地理解Condition对象了。它和wait()和notify()方法的作用是大致相同的。但是wait()和notify()方法是和synchronized关键字合作使用的，而Condition是与重入锁相关联的。通过Lock接口（重入锁就实现了这一接口）的Condition newCondition()方法可以生成一个与当前重入锁绑定的Condition实例。利用Condition对象，我们就可以让线程在合适的时间等待，或者在某一个特定的时刻得到通知，继续执行。

Condition接口提供的基本方法如下：

```
void await() throws InterruptedException;
void awaitUninterruptibly();
long awaitNanos(long nanosTimeout) throws InterruptedException;
boolean await(long time, TimeUnit unit) throws InterruptedException;
boolean awaitUntil(Date deadline) throws InterruptedException;
void signal();
void signalAll();
```

以上方法的含义如下：

- await()方法会使当前线程等待，同时释放当前锁，当其他线程中使用signal()或者signalAll()方法时，线程会重新获得锁并继续执行。或者当线程被中断时，也能跳出等待。这和Object.wait()方法很相似。
- awaitUninterruptibly()方法与await()方法基本相同，但是它并不会在等待过程中响应中断。
- signal()方法用于唤醒一个在等待中的线程。相对的signalAll()方法会唤醒所有在等待中的线程。这和Object.notify()方法很类似。

下面的代码简单地演示了Condition的功能：

```
01 public class ReenterLockCondition implements Runnable{
02     public static ReentrantLock lock=new ReentrantLock();
03     public static Condition condition = lock.newCondition();
04     @Override
05     public void run() {
06         try {
07             lock.lock();
08             condition.await();
09             System.out.println("Thread is going on");
10         } catch (InterruptedException e) {
11             e.printStackTrace();
12         }finally{
13             lock.unlock();
14         }
15     }
16     public static void main(String[] args) throws InterruptedException
17     {
18         ReenterLockCondition t1=new ReenterLockCondition();
19         Thread t1=new Thread(t1);
20         t1.start();
21         Thread.sleep(2000);
22         //通知线程t1继续执行
23         lock.lock();
24         condition.signal();
25         lock.unlock();
26     }
27 }
```

代码第3行，通过lock生成一个与之绑定的Condition对象。代码第8行，要求线程在Condition对象上进行等待。代码第23行，由主线程main发出通知，告知等待在Condition上的线程可以继续执行了。

和Object.wait()和notify()方法一样，当线程使用Condition.await()时，要求线程持有相关的重入锁，在Condition.await()调用后，这个线程会释放这把锁。同理，在Condition.signal()方法调用时，也要求线程先获得相关的锁。在signal()方法调用后，系统会从当前Condition对象的等待队列中，唤醒一个线程。一旦线程被唤醒，它会重新尝试获得与之绑定的重入锁，一旦成功获取，就可以继续执行了。因此，在signal()方法调用之后，一般需要释放相关的锁，谦让给被唤醒的线程，让它可以继续执行。比如，在本例中，第24行代码就释放了重入锁，如果省略第24行，那么，虽然已经唤醒了线程t1，但是由于它无法重新获得锁，因而也就无法真正的继续执行。

在JDK内部，重入锁和Condition对象被广泛地使用，以ArrayBlockingQueue为例（可以参阅“3.3 JDK并发容器”一节），它的put()方法实现如下：

```
//在ArrayBlockingQueue中的一些定义
private final ReentrantLock lock;
private final Condition notEmpty;
private final Condition notFull;
lock = new ReentrantLock(fair);
notEmpty = lock.newCondition();           //生成一个与lock绑定的C
notFull = lock.newCondition();
```

```

//put()方法的实现
public void put(E e) throws InterruptedException {
    if (e == null) throw new NullPointerException();
    final E[] items = this.items;
    final ReentrantLock lock = this.lock;
    lock.lockInterruptibly();           //对put()方法做同步
    try {
        try {
            while (count == items.length)    //如果当前队列已满
                notFull.await();           //等待队列有足够的空间
        } catch (InterruptedException ie) {
            notFull.signal();
            throw ie;
        }
        insert(e);                          //当notFull被通知时，该
    } finally {
        lock.unlock();
    }
}

private void insert(E x) {
    items[putIndex] = x;
    putIndex = inc(putIndex);
    ++count;
    notEmpty.signal();                   //通知需要take()的线程,
}

```

同理，对应take()方法实现如下：

```
public E take() throws InterruptedException {
    final ReentrantLock lock = this.lock;
    lock.lockInterruptibly();           //对take()方法做同步
    try {
        try {
            while (count == 0)          //如果队列为空
                notEmpty.await();        //则消费者队列要等待一个
        } catch (InterruptedException ie) {
            notEmpty.signal();
            throw ie;
        }
        E x = extract();
        return x;
    } finally {
        lock.unlock();
    }
}

private E extract() {
    final E[] items = this.items;
    E x = items[takeIndex];
    items[takeIndex] = null;
    takeIndex = inc(takeIndex);
    --count;
    notFull.signal();                   //通知put()线程队列已有
    return x;
}
```

```
}
```

3.1.3 允许多个线程同时访问：信号量（Semaphore）

信号量为多线程协作提供了更为强大的控制方法。广义上说，信号量是对锁的扩展。无论是内部锁`synchronized`还是重入锁`ReentrantLock`，一次都只允许一个线程访问一个资源，而信号量却可以指定多个线程，同时访问某一个资源。信号量主要提供了以下构造函数：

```
public Semaphore(int permits)
public Semaphore(int permits, boolean fair)    //第二个参数可以指定是否公平
```

在构造信号量对象时，必须要指定信号量的准入数，即同时能申请多少个许可。当每个线程每次只申请一个许可时，这就相当于指定了同时有多少个线程可以访问某一个资源。信号量的主要逻辑方法有：

```
public void acquire()
public void acquireUninterruptibly()
public boolean tryAcquire()
public boolean tryAcquire(long timeout, TimeUnit unit)
public void release()
```

`acquire()`方法尝试获得一个准入的许可。若无法获得，则线程会等待，直到有线程释放一个许可或者当前线程被中断。

`acquireUninterruptibly()`方法和`acquire()`方法类似，但是不响应中断。

tryAcquire()尝试获得一个许可，如果成功返回true，失败则返回false，它不会进行等待，立即返回。release()用于在线程访问资源结束后，释放一个许可，以使其他等待许可的线程可以进行资源访问。

在JDK的官方Javadoc中，就有一个有关信号量使用的简单实例，有兴趣的读者可以自行翻阅，这里我给出一个更加傻瓜化的例子：

```
01 public class SemapDemo implements Runnable{
02     final Semaphore semp = new Semaphore(5);
03     @Override
04     public void run() {
05         try {
06             semp.acquire();
07             //模拟耗时操作
08             Thread.sleep(2000);
09             System.out.println(Thread.currentThread().getId()+
10                 semp.release());
11         } catch (InterruptedException e) {
12             e.printStackTrace();
13         }
14     }
15
16     public static void main(String[] args) {
17         ExecutorService exec = Executors.newFixedThreadPool(20)
18         final SemapDemo demo=new SemapDemo();
19         for(int i=0;i<20;i++){
20             exec.submit(demo);
```



```
21      }  
22  }  
23 }
```

上述代码中，第7~9行为临界区管理代码，程序会限制执行这段代码的线程数。这里在第2行，申明了一个包含5个许可的信号量。这就意味着同时可以有5个线程进入代码段第7~9行。申请信号量使用`acquire()`操作，在离开时，务必使用`release()`释放信号量（代码第10行）。这和释放锁是一个道理。如果不幸发生了信号量的泄露（申请了但没有释放），那么可以进入临界区的线程数量就会越来越少，直到所有的线程均不可访问。在本例中，同时开启20个线程。观察这段程序的输出，你就会发现系统以5个线程一组为单位，依次输出带有线程ID的提示文本。

3.1.4 ReadWriteLock读写锁

`ReadWriteLock`是JDK5中提供的读写分离锁。读写分离锁可以有效地帮助减少锁竞争，以提升系统性能。用锁分离的机制来提升性能非常容易理解，比如线程A1、A2、A3进行写操作，B1、B2、B3进行读操作，如果使用重入锁或者内部锁，则理论上说所有读之间、读与写之间、写和写之间都是串行操作。当B1进行读取时，B2、B3则需要等待锁。由于读操作并不对数据的完整性造成破坏，这种等待显然是不合理。因此，读写锁就有了发挥功能的余地。

在这种情况下，读写锁允许多个线程同时读，使得B1、B2、B3之间真正并行。但是，考虑到数据完整性，写写操作和读写操作间依然是需要相互等待和持有锁的。总的来说，读写锁的访问约束如表3.1所

示。

表3.1 读写锁的访问约束情况

	读	写
读	非阻塞	阻塞
写	阻塞	阻塞

- 读-读不互斥：读读之间不阻塞。
- 读-写互斥：读阻塞写，写也会阻塞读。
- 写-写互斥：写写阻塞。

如果在系统中，读操作次数远远大于写操作，则读写锁就可以发挥最大的功效，提升系统的性能。这里我给出一个稍微夸张点的案例，来说明读写锁对性能的帮助。

```
01 public class ReadWriteLockDemo {
02     private static Lock lock=new ReentrantLock();
03     private static ReentrantReadWriteLock readWriteLock=new
ReentrantReadWriteLock();
04     private static Lock readLock = readWriteLock.readLock();
05     private static Lock writeLock = readWriteLock.writeLock();
06     private int value;
07
08     public Object handleRead(Lock lock) throws InterruptedException
09         try{
10             lock.lock();                //模拟读操作
```

```
11         Thread.sleep(1000);           //读操作的耗时越多，读写锁
12         return value;
13     }finally{
14         lock.unlock();
15     }
16 }
17
18 public void handleWrite(Lock lock,int index) throws Interr
19     try{
20         lock.lock();                   //模拟写操作
21         Thread.sleep(1000);
22         value=index;
23     }finally{
24         lock.unlock();
25     }
26 }
27
28 public static void main(String[] args) {
29     final ReadWriteLockDemo demo=new ReadWriteLockDemo();
30     Runnable readRunnale=new Runnable() {
31         @Override
32         public void run() {
33             try {
34                 demo.handleRead(readLock);
35 //                 demo.handleRead(lock);
36             } catch (InterruptedException e) {
37                 e.printStackTrace();
```

```

38         }
39     }
40 };
41     Runnable writeRunnale=new Runnable() {
42         @Override
43         public void run() {
44             try {
45                 demo.handleWrite(writeLock,new Random().ne
46 //                 demo.handleWrite(lock,new Random().nextI
47             } catch (InterruptedException e) {
48                 e.printStackTrace();
49             }
50         }
51     };
52
53     for(int i=0;i<18;i++){
54         new Thread(readRunnale).start();
55     }
56
57     for(int i=18;i<20;i++){
58         new Thread(writeRunnale).start();
59     }
60 }
61 }

```

上述代码中，第11行和第21行分别模拟了一个非常耗时的操作，让线程耗时1秒钟。它们分别对应读耗时和写耗时。代码第34和45行，分

别是读线程和写线程。在这里，第34行使用读锁，第35行使用写锁。第53~55行开启18个读线程，第57~59行，开启两个写线程。由于这里使用了读写分离，因此，读线程完全并行，而写会阻塞读，因此，实际上这段代码运行大约2秒多就能结束（写线程之间是实际串行的）。而如果使用第35行代替第34行，使用第46行代替第45行执行上述代码，即，使用普通的重入锁代替读写锁。那么所有的读和写线程之间都必须相互等待，因此整个程序的执行时间将长达20余秒。

3.1.5 倒计时器：CountDownLatch

CountDownLatch是一个非常实用的多线程控制工具类。“Count Down”在英文中意为倒数，Latch为门闩的意思。如果翻译成为倒数门闩，我想大家都会觉得不知所云吧！因此，这里简单地称之为倒数计数器。在这里，门闩的含义是：把门锁起来，不让里面的线程跑出来。因此，这个工具通常用来控制线程等待，它可以让某一个线程等待直到倒计时结束，再开始执行。

对于倒计时器，一种典型的场景就是火箭发射。在火箭发射前，为了保证万无一失，往往还要进行各项设备、仪器的检查。只有等所有的检查完毕后，引擎才能点火。这种场景就非常适合使用CountDownLatch。它可以使得点火线程等待所有检查线程全部完工后，再执行。

CountDownLatch的构造函数接收一个整数作为参数，即当前这个计数器的计数个数。

```
public CountDownLatch(int count)
```

下面这个简单的示例，演示了CountDownLatch的使用。

```
01 public class CountDownLatchDemo implements Runnable {
02     static final CountDownLatch end = new CountDownLatch(10);
03     static final CountDownLatchDemo demo=new CountDownLatchDem
04     @Override
05     public void run() {
06         try {
07             //模拟检查任务
08             Thread.sleep(new Random().nextInt(10)*1000);
09             System.out.println("check complete");
10             end.countDown();
11         } catch (InterruptedException e) {
12             e.printStackTrace();
13         }
14     }
15     public static void main(String[] args) throws InterruptedE
16         ExecutorService exec = Executors.newFixedThreadPool(10
17         for(int i=0;i<10;i++){
18             exec.submit(demo);
19         }
20         //等待检查
21         end.await();
22         //发射火箭
23         System.out.println("Fire!");
24         exec.shutdown();
25     }
```

上述代码第2行，生成一个CountDownLatch实例。计数数量为10。这表示需要有10个线程完成任务，等待在CountDownLatch上的线程才能继续执行。代码第10行，使用了CountDownLatch.countdown()方法，也就是通知CountDownLatch，一个线程已经完成了任务，倒计时器可以减1啦。第21行，使用CountDownLatch.await()方法，要求主线程等待所有10个检查任务全部完成。待10个任务全部完成后，主线程才能继续执行。

上述案例的执行逻辑可以用图3.1简单表示。

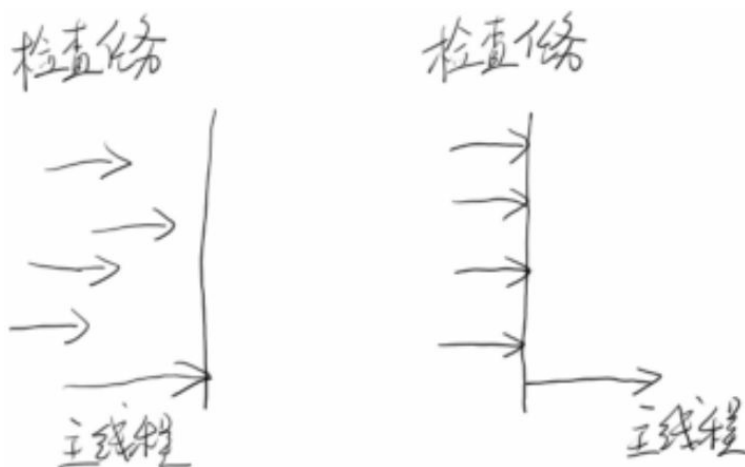


图3.1 CountDownLatch示意图

主线程在CountDownLatch上等待，当所有检查任务全部完成后，主线程方能继续执行。

3.1.6 循环栅栏：CyclicBarrier

CyclicBarrier是另外一种多线程并发控制实用工具。和

CountDownLatch非常类似，它也可以实现线程间的计数等待，但它的功能比CountDownLatch更加复杂且强大。

CyclicBarrier可以理解为循环栅栏。栅栏就是一种障碍物，比如，通常在私人宅邸的周围就可以围上一圈栅栏，阻止闲杂人等入内。这里当然就是用来阻止线程继续执行，要求线程在栅栏处等待。前面Cyclic意为循环，也就是说这个计数器可以反复使用。比如，假设我们将计数器设置为10，那么凑齐第一批10个线程后，计数器就会归零，然后接着凑齐下一批10个线程，这就是循环栅栏内在的含义。

CyclicBarrier的使用场景也很丰富。比如，司令下达命令，要求10个士兵一起去完成一项任务。这时，就会要求10个士兵先集合报道，接着，一起雄赳赳气昂昂地去执行任务。当10个士兵把自己手头的任务都执行完成了，那么司令才能对外宣布，任务完成！

比CountDownLatch略微强大一些，CyclicBarrier可以接收一个参数作为barrierAction。所谓barrierAction就是当计数器一次计数完成后，系统会执行的动作。如下构造函数，其中，parties表示计数总数，也就是参与的线程总数。

```
public CyclicBarrier(int parties, Runnable barrierAction)
```

下面的示例使用CyclicBarrier演示了上述司令命令士兵完成任务的场景。

```
01 public class CyclicBarrierDemo {
02     public static class Soldier implements Runnable {
03         private String soldier;
04         private final CyclicBarrier cyclic;
```



```
05
06     Soldier(CyclicBarrier cyclic, String soldierName) {
07         this.cyclic = cyclic;
08         this.soldier = soldierName;
09     }
10
11     public void run() {
12         try {
13             //等待所有士兵到齐
14             cyclic.await();
15             doWork();
16             //等待所有士兵完成工作
17             cyclic.await();
18         } catch (InterruptedException e) {
19             e.printStackTrace();
20         } catch (BrokenBarrierException e) {
21             e.printStackTrace();
22         }
23     }
24
25     void doWork() {
26         try {
27             Thread.sleep(Math.abs(new Random().nextInt())%1
28         } catch (InterruptedException e) {
29             e.printStackTrace();
30         }
31         System.out.println(soldier + ":任务完成");
```

```
32     }
33 }
34
35 public static class BarrierRun implements Runnable {
36     boolean flag;
37     int N;
38     public BarrierRun(boolean flag, int N) {
39         this.flag = flag;
40         this.N = N;
41     }
42
43     public void run() {
44         if (flag) {
45             System.out.println("司令:[士兵" + N + "个, 任务完
46         } else {
47             System.out.println("司令:[士兵" + N + "个, 集合完
48             flag = true;
49         }
50     }
51 }
52
53 public static void main(String args[]) throws InterruptedException
54     final int N = 10;
55     Thread[] allSoldier=new Thread[N];
56     boolean flag = false;
57     CyclicBarrier cyclic = new CyclicBarrier(N, new Barrie
58     //设置屏障点, 主要是为了执行这个方法
```

```

59         System.out.println("集合队伍! ");
60         for (int i = 0; i < N; ++i) {
61             System.out.println("士兵 "+i+" 报道!");
62             allSoldier[i]=new Thread(new Soldier(cyclic, "士兵
63             allSoldier[i].start();
64         }
65     }
66 }

```

上述代码第57行，创建了CyclicBarrier实例，并将计数器设置为10，并要求在计数器达到指标时，执行第43行的run()方法。每一个士兵线程会执行第11行定义的run()方法。在第14行，每一个士兵线程都会等待，直到所有的士兵都集合完毕。集合完毕后，意味着CyclicBarrier的一次计数完成，当再一次调用CyclicBarrier.await()时，会进行下一次计数。第15行，模拟了士兵的任务。当一个士兵任务执行完毕后，他就会要求CyclicBarrier开始下一次计数，这次计数主要目的是监控是否所有的士兵都已经完成了任务。一旦任务全部完成，第35行定义的BarrierRun就会被调用，打印相关信息。

上述代码的执行输出如下：

```

集合队伍！
士兵 0 报道！
//篇幅有限，省略其他几个士兵
士兵 9 报道！
司令:[士兵10个，集合完毕！]
士兵 0:任务完成
//篇幅有限，省略其他几个士兵

```

士兵 4:任务完成

司令:[士兵10个,任务完成!]

整个工作过程的图示如图3.2所示。

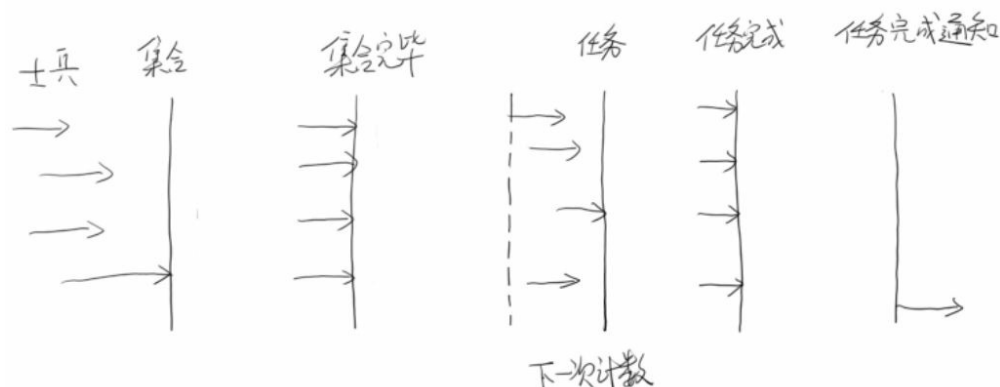


图3.2 CyclicBarrier工作示意图

CyclicBarrier.await()方法可能会抛出两个异常。一个是InterruptedException，也就是在等待过程中，线程被中断，应该说这是一个非常通用的异常。大部分迫使线程等待的方法都可能会抛出这个异常，使得线程在等待时依然可以响应外部紧急事件。另外一个异常则是CyclicBarrier特有的BrokenBarrierException。一旦遇到这个异常，则表示当前的CyclicBarrier已经破损了，可能系统已经没有办法等待所有线程到齐了。如果继续等待，可能就是徒劳无功的，因此，还是就地散货，打道回府吧！上述代码第18~22行处理了这2种异常。

如果我们在上述代码的第63行后，插入以下代码，使得第5个士兵线程产生中断：

```
if(i==5){
    allSoldier[0].interrupt();
}
```

如果这样做，我们很可能就会得到1个InterruptedException和9个BrokenBarrierException。这个InterruptedException就是被中断线程抛出的。而其他9个BrokenBarrierException，则是等待在当前CyclicBarrier上的线程抛出的。这个异常可以避免其他9个线程进行永久的、无谓的等待（因为其中一个线程已经被中断，等待是没有结果的）。

3.1.7 线程阻塞工具类：LockSupport

LockSupport是一个非常方便实用的线程阻塞工具，它可以在线程内任意位置让线程阻塞。和Thread.suspend()相比，它弥补了由于resume()在前发生，导致线程无法继续执行的情况。和Object.wait()相比，它不需要先获得某个对象的锁，也不会抛出InterruptedException异常。

LockSupport的静态方法park()可以阻塞当前线程，类似的还有parkNanos()、parkUntil()等方法。它们实现了一个限时的等待。

大家应该还记得，我们在第2章中提到的那个有关suspend()永久卡死线程的例子吧！现在，我们可以用LockSupport重写这个程序：

```
01 public class LockSupportDemo {
02     public static Object u = new Object();
03     static ChangeObjectThread t1 = new ChangeObjectThread("t1"
04     static ChangeObjectThread t2 = new ChangeObjectThread("t2"
05
06     public static class ChangeObjectThread extends Thread {
07         public ChangeObjectThread(String name){
```

```

08         super.setName(name);
09     }
10     @Override
11     public void run() {
12         synchronized (u) {
13             System.out.println("in "+getName());
14             LockSupport.park();
15         }
16     }
17 }
18
19 public static void main(String[] args) throws InterruptedException
20     {
21         t1.start();
22         Thread.sleep(100);
23         t2.start();
24         LockSupport.unpark(t1);
25         LockSupport.unpark(t2);
26         t1.join();
27         t2.join();
28     }

```

注意，这里只是将原来的suspend()和resume()方法用park()和unpark()方法做了替换。当然，我们依然无法保证unpark()方法发生在park()方法之后。但是执行这段代码，你会发现，它自始至终都可以正常的结束，不会因为park()方法而导致线程永久性的挂起。

这是因为LockSupport类使用类似信号量的机制。它为每一个线程准备了一个许可，如果许可可用，那么park()函数会立即返回，并且消费这个许可（也就是将许可变为不可用），如果许可不可用，就会阻塞。而unpark()则使得一个许可变为可用（但是和信号量不同的是，许可不能累加，你不可能拥有超过一个许可，它永远只有一个）。

这个特点使得：即使unpark()操作发生在park()之前，它也可以使下一次的park()操作立即返回。这也就是上述代码可顺利结束的主要原因。

同时，处于park()挂起状态的线程不会像suspend()那样还给出一个令人费解的Runnable的状态。它会非常明确地给出一个WAITING状态，甚至还会标注是park()引起的：

```
"t1" #8 prio=5 os_prio=0 tid=0x00b1a400 nid=0x1994 waiting on con
  java.lang.Thread.State: WAITING (parking)
    at sun.misc.Unsafe.park(Native Method)
    at java.util.concurrent.locks.LockSupport.park(LockSupport.
    at geym.conc.ch3.ls.LockSupportDemo$ChangeObjectThread.ru
    - locked <0x048b2680> (a java.lang.Object)
```

这使得分析问题格外方便。此外，如果你使用park(Object)函数，还可以为当前线程设置一个阻塞对象。这个阻塞对象会出现在线程Dump中。这样在分析问题时，就更加方便了。

比如，如果我们将上述代码第14行的park()改为：

```
LockSupport.park(this);
```

那么在线程Dump时，你可能会看到如下信息：

```
"t1" #8 prio=5 os_prio=0 tid=0x0117ac00 nid=0x2034 waiting on con
  java.lang.Thread.State: WAITING (parking)
    at sun.misc.Unsafe.park(Native Method)
      - parking to wait for  <0x048b4738> (a geym.conc.ch3.ls.
Demo$ChangeObjectThread)
    at java.util.concurrent.locks.LockSupport.park(LockSupport
    at geym.conc.ch3.ls.LockSupportDemo$ChangeObjectThread.ru
(LockSupportDemo.java:18)
      - locked <0x048b2808> (a java.lang.Object)
```

注意，在堆栈中，我们甚至还看到了当前线程等待的对象，这里就是ChangeObjectThread实例。

除了有定时阻塞的功能外，LockSupport.park()还能支持中断影响。但是和其他接收中断的函数很不一样，LockSupport.park()不会抛出InterruptedException异常。它只是会默默的返回，但是我们可以从Thread.interrupted()等方法获得中断标记。

```
01 public class LockSupportIntDemo {
02     public static Object u = new Object();
03     static ChangeObjectThread t1 = new ChangeObjectThread("t1"
04     static ChangeObjectThread t2 = new ChangeObjectThread("t2"
05
06     public static class ChangeObjectThread extends Thread {
07         public ChangeObjectThread(String name){
08             super.setName(name);
```



```
09     }
10     @Override
11     public void run() {
12         synchronized (u) {
13             System.out.println("in "+getName());
14             LockSupport.park();
15             if(Thread.interrupted()){
16                 System.out.println(getName()+" 被中断了");
17             }
18         }
19         System.out.println(getName()+"执行结束");
20     }
21 }
22
23 public static void main(String[] args) throws InterruptedException
24     t1.start();
25     Thread.sleep(100);
26     t2.start();
27     t1.interrupt();
28     LockSupport.unpark(t2);
29 }
30 }
```

注意上述代码在第27行，中断了处于park()状态的t1。之后，t1可以马上响应这个中断，并且返回。之后在外面等待的t2才可以进入临界区，并最终由LockSupport.unpark(t2)操作使其运行结束。

in t1

t1 被中断了

t1 执行结束

in t2

t2执行结束

3.2 线程复用：线程池

多线程的软件设计方法确实可以最大限度地发挥现代多核处理器的计算能力，提高生产系统的吞吐量和性能。但是，若不加控制和管理地随意使用线程，对系统的性能反而会产生不利的影响。

一种最为简单的线程创建和回收的方法类似如下代码：

```
new Thread(new Runnable(){
    @Override
    public void run() {
        //do sth.
    }
}).start();
```

以上代码创建了一个线程，并在run()方法结束后，自动回收该线程。在简单的应用系统中，这段代码并没有太多问题。但是在真实的生产环境中，系统由于真实环境的需要，可能会开启很多线程来支撑其应用。而当线程数量过大时，反而会耗尽CPU和内存资源。

首先，虽然与进程相比，线程是一种轻量级的工具，但其创建和关闭依然需要花费时间，如果为每一个小的任务都创建一个线程，很有可能出现创建和销毁线程所占用的时间大于该线程真实工作所消耗的时间的情况，反而会得不偿失。

其次，线程本身也是要占用内存空间的，大量的线程会抢占宝贵的内存资源，如果处理不当，可能会导致Out of Memory异常。即便没

有，大量的线程回收也会给GC带来很大的压力，延长GC的停顿时间。

因此，对线程的使用必须掌握一个度，在有限的范围内，增加线程的数量可以明显提高系统的吞吐量，但一旦超出了这个范围，大量的线程只会拖垮应用系统。因此，在生产环境中使用线程，必须对其加以控制和管理。

注意：在实际生产环境中，线程的数量必须得到控制。盲目的大量创建线程对系统性能是有伤害的。

3.2.1 什么是线程池

为了避免系统频繁地创建和销毁线程，我们可以让创建的线程进行复用。如果大家进行过数据库开发，对数据库连接池应该不会陌生。为了避免每次数据库查询都重新建立和销毁数据库连接，我们可以使用数据库连接池维护一些数据库连接，让他们长期保持在一个激活状态。当系统需要使用数据库时，并不是创建一个新的连接，而是从连接池中获得一个可用的连接即可。反之，当需要关闭连接时，并不真的把连接关闭，而是将这个连接“还”给连接池即可。通过这种方式，可以节约不少创建和销毁对象的时间。

线程池也是类似的概念。线程池中，总有那么几个活跃线程。当你需要使用线程时，可以从池子中随便拿一个空闲线程，当完成工作时，并不急着关闭线程，而是将这个线程退回到池子，方便其他人使用。

简而言之，在使用线程池后，创建线程变成了从线程池获得空闲线程，关闭线程变成了向池子归还线程，如图3.3所示。

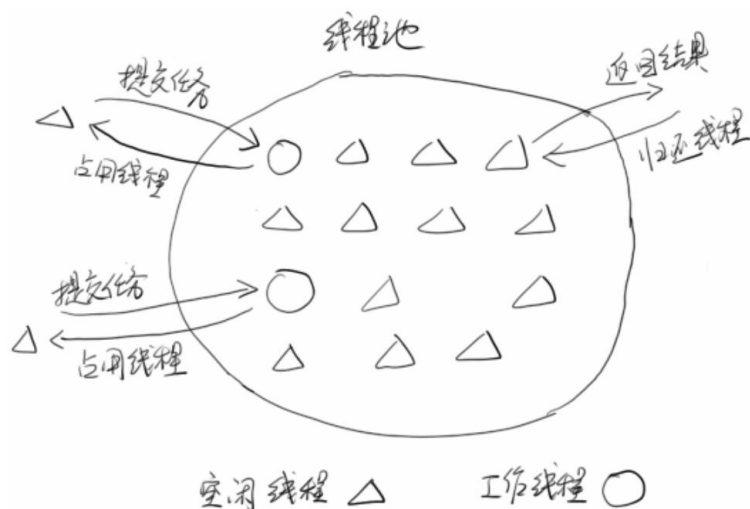


图3-3 线程池的作用

3.2.2 不要重复发明轮子：JDK对线程池的支持

为了能够更好地控制多线程，JDK提供了一套Executor框架，帮助开发人员有效地进行线程控制，其本质就是一个线程池。它的核心成员如图3.4所示。

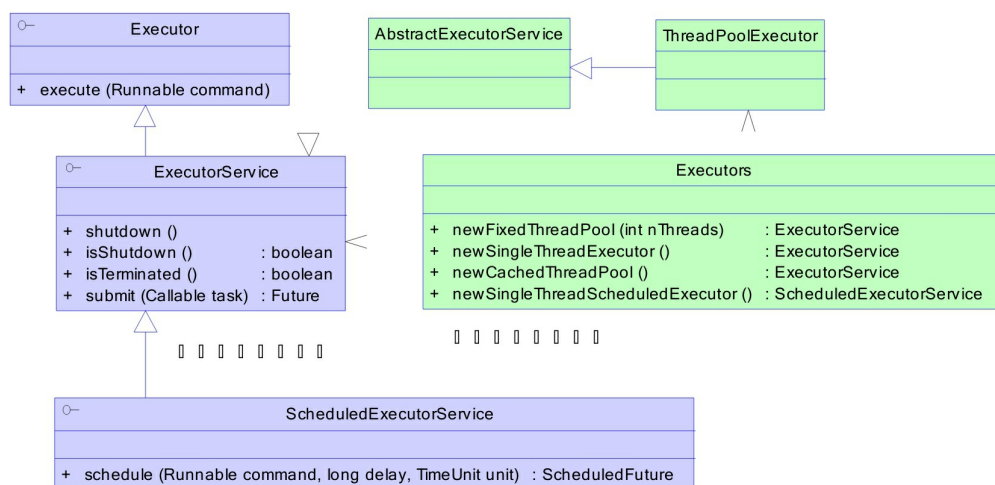


图3-4 Executor框架结构图

以上成员均在java.util.concurrent包中，是JDK并发包的核心类。其中ThreadPoolExecutor表示一个线程池。Executors类则扮演着线程池工厂的角色，通过Executors可以取得一个拥有特定功能的线程池。从UML图中亦可知，ThreadPoolExecutor类实现了Executor接口，因此通过这个接口，任何Runnable的对象都可以被ThreadPoolExecutor线程池调度。

Executor框架提供了各种类型的线程池，主要有以下工厂方法：

```
public static ExecutorService newFixedThreadPool(int nThreads)
public static ExecutorService newSingleThreadExecutor()
public static ExecutorService newCachedThreadPool()
public static ScheduledExecutorService newSingleThreadScheduledEx
public static ScheduledExecutorService newScheduledThreadPool(int
```

以上工厂方法分别返回具有不同工作特性的线程池。这些线程池工厂方法的具体说明如下。

- **newFixedThreadPool()方法：**该方法返回一个固定线程数量的线程池。该线程池中的线程数量始终不变。当有一个新的任务提交时，线程池中若有空闲线程，则立即执行。若没有，则新的任务会被暂存在一个任务队列中，待有线程空闲时，便处理在任务队列中的任务。
- **newSingleThreadExecutor()方法：**该方法返回一个只有一个线程的线程池。若多余一个任务被提交到该线程池，任务会被保存在一个任务队列中，待线程空闲，按先入先出的顺序执行队列中的任务。
- **newCachedThreadPool()方法：**该方法返回一个可根据实际情况调

整线程数量的线程池。线程池的线程数量不确定，但若有空闲线程可以复用，则会优先使用可复用的线程。若所有线程均在工作，又有新的任务提交，则会创建新的线程处理任务。所有线程在当前任务执行完毕后，将返回线程池进行复用。

- `newSingleThreadScheduledExecutor()`方法：该方法返回一个 `ScheduledExecutorService` 对象，线程池大小为1。
`ScheduledExecutorService` 接口在 `ExecutorService` 接口之上扩展了在给定时间执行某任务的功能，如在某个固定的延时之后执行，或者周期性执行某个任务。
- `newScheduledThreadPool()`方法：该方法也返回一个 `ScheduledExecutorService` 对象，但该线程池可以指定线程数量。

1. 固定大小的线程池

这里，我们以 `newFixedThreadPool()` 为例，简单地展示线程池的使用：

```
01 public class ThreadPoolDemo {
02     public static class MyTask implements Runnable {
03         @Override
04         public void run() {
05             System.out.println(System.currentTimeMillis() + ":
06                 + Thread.currentThread().getId());
07             try {
08                 Thread.sleep(1000);
09             } catch (InterruptedException e) {
10                 e.printStackTrace();
11             }
12     }
```

```
12     }
13 }
14
15 public static void main(String[] args) {
16     MyTask task = new MyTask();
17     ExecutorService es = Executors.newFixedThreadPool(5);
18     for (int i = 0; i < 10; i++) {
19         es.submit(task);
20     }
21 }
22 }
```

上述代码中，第17行创建了固定大小的线程池，内有5个线程。在第19行，依次向线程池提交了10个任务。此后，线程池就会安排调度这10个任务。每个任务都会将自己的执行时间和执行这个线程的ID打印出来，并且在这里，安排每个任务要执行1秒钟。

执行上述代码，可以得到类似以下输出：

```
1426510293450:Thread ID:8
1426510293450:Thread ID:9
1426510293450:Thread ID:12
1426510293450:Thread ID:10
1426510293450:Thread ID:11
1426510294450:Thread ID:12
1426510294450:Thread ID:11
1426510294450:Thread ID:8
1426510294450:Thread ID:10
```


1426510294450:Thread ID:9

这个输出就表示这10个线程的执行情况。很显然，前5个任务和后5个任务的执行时间正好相差1秒钟（注意时间戳的单位是毫秒），并且前5个任务的线程ID和后5个任务也是完全一致的（都是8、9、10、11、12）。这说明在这10个任务中，是分成2批次执行的。这也完全符合一个只有5个线程的线程池的行为。

有兴趣的读者可以将其改造成`newCachedThreadPool()`，看看任务的分配情况会有何变化？

2. 计划任务

另外一个值得注意的方法是`newScheduledThreadPool()`。它返回一个`ScheduledExecutorService`对象，可以根据时间需要对线程进行调度。它的一些主要方法如下：

```
public ScheduledFuture<?> schedule(Runnable command, long delay,
public ScheduledFuture<?> scheduleAtFixedRate(Runnable command,
                                                long initialDelay,
                                                long period,
                                                TimeUnit unit);
public ScheduledFuture<?> scheduleWithFixedDelay(Runnable command,
                                                long initialDelay,
                                                long delay,
                                                TimeUnit unit);
```

与其他几个线程池不同，`ScheduledExecutorService`并不一定会立即安排执行任务。它其实是起到了计划任务的作用。它会在指定的时间，

对任务进行调度。如果大家使用过Linux下的crontab工具应该就能很容易地理解它了。

作为说明，这里给出了三个方法。方法schedule()会在给定时间，对任务进行一次调度。方法scheduleAtFixedRate()和scheduleWithFixedDelay()会对任务进行周期性的调度。但是两者有一点小小的区别，如图3.5所示。

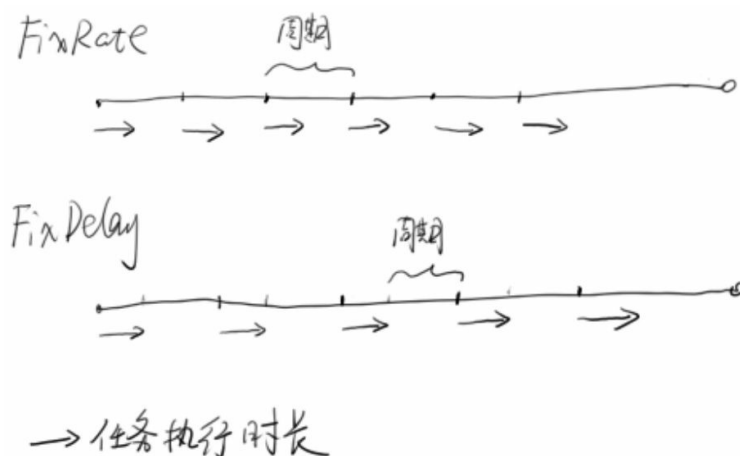


图3.5 FixedRate和FixDelay区别

对于FixedRate方式来说，任务调度的频率是一定的。它是以上一个任务开始执行时间为起点，之后的period时间，调度下一次任务。而FixDelay则是在上一个任务结束后，再经过delay时间进行任务调度。

由于担心我的解释不够周全，我也很乐意将官方文档中的描述贴出来供大家参考，从而可以更精确地理解两者的差别：

- scheduleAtFixedRate
 - Creates and executes a periodic action that becomes enabled first after the given initial delay, and subsequently with the given period; that is executions will commence after initialDelay then

initialDelay+period, then initialDelay + 2 * period, and so on.

- 翻译：创建一个周期性任务。任务开始于给定的初始延时。后续的任务按照给定的周期进行：后续第一个任务将会在 initialDelay+period时执行，后续第二个任务将在 initialDelay+2*period时进行，依此类推。

- scheduleWithFixedDelay

- Creates and executes a periodic action that becomes enabled first after the given initial delay, and subsequently with the given delay between the termination of one execution and the commencement of the next.
- 翻译：创建并执行一个周期性任务。任务开始于初始延时时间，后续任务将会按照给定的延时进行，即上一个任务的结束时间到下一个任务的开始时间的时间差。

下面的例子使用scheduleAtFixedRate()方法调度一个任务。这个任务会执行1秒钟时间，调度周期是2秒。也就是说每2秒钟，任务就会被执行一次。

```
01 public class ScheduledExecutorServiceDemo {
02     public static void main(String[] args) {
03         ScheduledExecutorService ses=Executors.newScheduledThr
04         //如果前面的任务没有完成，则调度也不会启动
05         ses.scheduleAtFixedRate(new Runnable() {
06             @Override
07             public void run() {
08                 try {
```

```
09             Thread.sleep(1000);
10             System.out.println(System.currentTimeMillis());
11         } catch (InterruptedException e) {
12             e.printStackTrace();
13         }
14     }
15     }, 0, 2, TimeUnit.SECONDS);
16 }
17 }
```

执行上述代码，一种输出的可能如下：

```
1426515345
1426515347
1426515349
1426515351
```

上述输出的单位是秒。可以看到，时间间隔是2秒。

这里还想说一个有意思的事情，如果任务的执行时间超过调度时间，会发生什么情况呢？比如，这里调度周期是2秒，如果任务的执行时间是8秒，是不是会出现多个任务堆叠在一起呢？

实际上，`ScheduledExecutorService`不会让任务堆叠出现。我们将第9行的代码改为：

```
Thread.sleep(8000);
```

再执行上述代码，你就会发现任务的执行周期不再是2秒，而是变

成了8秒。如下所示，是一种可能的结果。

```
1426516323
1426516331
1426516339
1426516347
1426516355
```

也就是说，周期如果太短，那么任务就会在上一个任务结束后，立即被调用。可以想象，如果采用`scheduleWithFixedDelay()`，并且按照修改8秒，调度周期2秒计，那么任务的实际间隔将是10秒，大家可以自行尝试。

另外一个值得注意的问题是，调度程序实际上并不保证任务会无限期的持续调用。如果任务本身抛出了异常，那么后续的所有执行都会被中断，因此，如果你想让你的任务持续稳定的执行，那么做好异常处理就非常重要，否则，你很有可能观察到你的调度器无疾而终。

注意：如果任务遇到异常，那么后续的所有子任务都会停止调度，因此，必须保证异常被及时处理，为周期性任务的稳定调度提供条件。

3.2.3 刨根究底：核心线程池的内部实现

对于核心的几个线程池，无论是`newFixedThreadPool()`方法、`newSingleThreadExecutor()`还是`newCachedThreadPool()`方法，虽然看起

来创建的线程有着完全不同的功能特点，但其内部实现均使用了ThreadPoolExecutor实现。下面给出了这三个线程池的实现方式：

```
public static ExecutorService newFixedThreadPool(int nThreads) {  
    return new ThreadPoolExecutor(nThreads, nThreads,  
                                    0L, TimeUnit.MILLISECONDS,  
                                    new LinkedBlockingQueue<Runnable>())  
}  
  
public static ExecutorService newSingleThreadExecutor() {  
    return new FinalizableDelegatedExecutorService  
        (new ThreadPoolExecutor(1, 1,  
                                   0L, TimeUnit.MILLISECONDS,  
                                   new LinkedBlockingQueue<Runnable>())  
)  
}  
  
public static ExecutorService newCachedThreadPool() {  
    return new ThreadPoolExecutor(0, Integer.MAX_VALUE,  
                                    60L, TimeUnit.SECONDS,  
                                    new SynchronousQueue<Runnable>())  
}
```

由以上线程池的实现代码可以看到，它们都只是ThreadPoolExecutor类的封装。为何ThreadPoolExecutor有如此强大的功能呢？来看一下ThreadPoolExecutor最重要的构造函数：

```
public ThreadPoolExecutor(int corePoolSize,  
                          int maximumPoolSize,
```

```
long keepAliveTime,  
TimeUnit unit,  
BlockingQueue<Runnable> workQueue,  
ThreadFactory threadFactory,  
RejectedExecutionHandler handler)
```

函数的参数含义如下。

- **corePoolSize**: 指定了线程池中的线程数量。
- **maximumPoolSize**: 指定了线程池中的最大线程数量。
- **keepAliveTime**: 当线程池线程数量超过**corePoolSize**时，多余的空闲线程的存活时间。即，超过**corePoolSize**的空闲线程，在多长时间之内，会被销毁。
- **unit**: **keepAliveTime**的单位。
- **workQueue**: 任务队列，被提交但尚未被执行的任务。
- **threadFactory**: 线程工厂，用于创建线程，一般用默认的即可。
- **handler**: 拒绝策略。当任务太多来不及处理，如何拒绝任务。

以上参数中，大部分都很简单，只有**workQueue**和**handler**需要进行详细说明。

参数**workQueue**指被提交但未执行的任务队列，它是一个**BlockingQueue**接口的对象，仅用于存放**Runnable**对象。根据队列功能分类，在**ThreadPoolExecutor**的构造函数中可使用以下几种**BlockingQueue**。

- 直接提交的队列：该功能由**SynchronousQueue**对象提供。
SynchronousQueue是一个特殊的**BlockingQueue**。

`SynchronousQueue`没有容量，每一个插入操作都要等待一个相应的删除操作，反之，每一个删除操作都要等待对应的插入操作。如果使用`SynchronousQueue`，提交的任务不会被真实的保存，而总是将新任务提交给线程执行，如果没有空闲的进程，则尝试创建新的进程，如果进程数量已经达到最大值，则执行拒绝策略。因此，使用`SynchronousQueue`队列，通常要设置很大的`maximumPoolSize`值，否则很容易执行拒绝策略。

- 有界的任务队列：有界的任务队列可以使用`ArrayBlockingQueue`实现。`ArrayBlockingQueue`的构造函数必须带一个容量参数，表示该队列的最大容量，如下所示。

```
public ArrayBlockingQueue(int capacity)
```

当使用有界的任务队列时，若有新的任务需要执行，如果线程池的实际线程数小于`corePoolSize`，则会优先创建新的线程，若大于`corePoolSize`，则会将新任务加入等待队列。若等待队列已满，无法加入，则在总线程数不大于`maximumPoolSize`的前提下，创建新的进程执行任务。若大于`maximumPoolSize`，则执行拒绝策略。可见，有界队列仅当在任务队列装满时，才可能将线程数提升到`corePoolSize`以上，换言之，除非系统非常繁忙，否则确保核心线程数维持在在`corePoolSize`。

- 无界的任务队列：无界任务队列可以通过`LinkedBlockingQueue`类实现。与有界队列相比，除非系统资源耗尽，否则无界的任务队列不存在任务入队失败的情况。当有新的任务到来，系统的线程数小于`corePoolSize`时，线程池会生成新的线程执行任务，但当系统的线程数达到`corePoolSize`后，就不会继续增加。若后续仍有新的任务加入，而又没有空闲的线程资源，则任务直接进入队

列等待。若任务创建和处理的速度差异很大，无界队列会保持快速增长，直到耗尽系统内存。

- 优先任务队列：优先任务队列是带有执行优先级的队列。它通过PriorityBlockingQueue实现，可以控制任务的执行先后顺序。它是一个特殊的无界队列。无论是有界队列ArrayBlockingQueue，还是未指定大小的无界队列LinkedBlockingQueue都是按照先进先出算法处理任务的。而PriorityBlockingQueue则可以根据任务自身的优先级顺序先后执行，在确保系统性能的同时，也能有很好的质量保证（总是确保高优先级的任务先执行）。

回顾newFixedThreadPool()方法的实现。它返回了一个corePoolSize和maximumPoolSize大小一样的，并且使用了LinkedBlockingQueue任务队列的线程池。因为对于固定大小的线程池而言，不存在线程数量的动态变化，因此corePoolSize和maximumPoolSize可以相等。同时，它使用无界队列存放无法立即执行的任务，当任务提交非常频繁的时候，该队列可能迅速膨胀，从而耗尽系统资源。

newSingleThreadExecutor()返回的单线程线程池，是newFixedThreadPool()方法的一种退化，只是简单的将线程池线程数量设置为1。

newCachedThreadPool()方法返回corePoolSize为0，maximumPoolSize无穷大的线程池，这意味着在没有任务时，该线程池内无线程，而当任务被提交时，该线程池会使用空闲的线程执行任务，若无空闲线程，则将任务加入SynchronousQueue队列，而SynchronousQueue队列是一种直接提交的队列，它总会迫使线程池增加新的线程执行任务。当任务执行完毕后，由于corePoolSize为0，因此空闲线程又会在指定时间内（60秒）被回收。

对于newCachedThreadPool(), 如果同时有大量任务被提交, 而任务的执行又不那么快时, 那么系统便会开启等量的线程处理, 这样做法可能会很快耗尽系统的资源。

注意: 使用自定义线程池时, 要根据应用的具体情况, 选择合适的并发队列作为任务的缓冲。当线程资源紧张时, 不同的并发队列对系统行为和性能的影响均不同。

这里给出ThreadPoolExecutor线程池的核心调度代码, 这段代码也充分体现了上述线程池的工作逻辑:

```
01 public void execute(Runnable command) {
02     if (command == null)
03         throw new NullPointerException();
04     int c = ctl.get();
05     if (workerCountOf(c) < corePoolSize) {
06         if (addWorker(command, true))
07             return;
08         c = ctl.get();
09     }
10     if (isRunning(c) && workQueue.offer(command)) {
11         int recheck = ctl.get();
12         if (! isRunning(recheck) && remove(command))
13             reject(command);
14         else if (workerCountOf(recheck) == 0)
15             addWorker(null, false);
16     }
17     else if (!addWorker(command, false))
```

```
18         reject(command);
19     }
```

代码第5行的`workerCountOf()`函数取得了当前线程池的线程总数。当线程总数小于`corePoolSize`核心线程数时，会将任务通过`addWorker()`方法直接调度执行。否则，则在第10行代码处（`workQueue.offer()`）进入等待队列。如果进入等待队列失败（比如有界队列到达了上限，或者使用了`SynchronousQueue`），则会执行第17行，将任务直接提交给线程池。如果当前线程数已经达到`maximumPoolSize`，则提交失败，就执行第18行的拒绝策略。

调度逻辑可以总结为如图3.6所示。

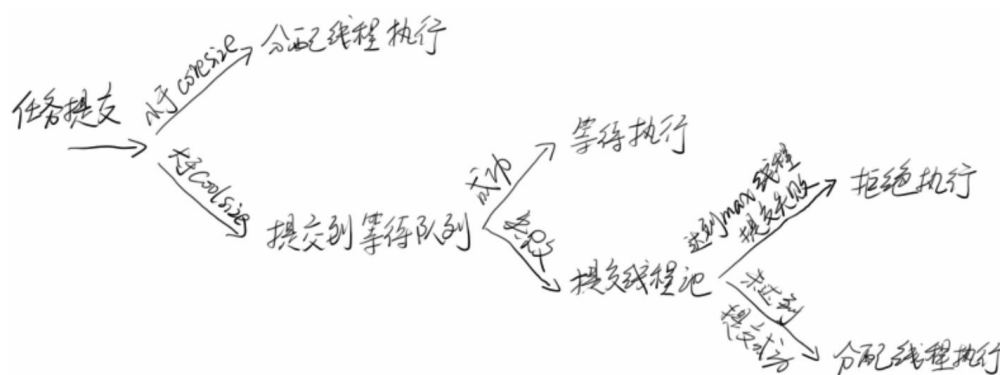


图3.6 ThreadPoolExecutor的任务调度逻辑

3.2.4 超负载了怎么办：拒绝策略

ThreadPoolExecutor的最后一个参数指定了拒绝策略。也就是当任务数量超过系统实际承载能力时，该如何处理呢？这时就要用到拒绝策略了。拒绝策略可以说是系统超负荷运行时的补救措施，通常由于压力太大而引起的，也就是线程池中的线程已经用完了，无法继续为新任务

服务，同时，等待队列中也已经排满了，再也塞不下新任务了。这时，我们就需要有一套机制，合理地处理这个问题。

JDK内置提供了四种拒绝策略，如图3.7所示。

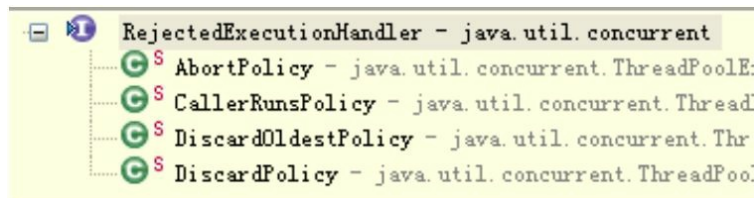


图3.7 JDK内置的拒绝策略

JDK内置的拒绝策略如下。

- **AbortPolicy策略：**该策略会直接抛出异常，阻止系统正常工作。
- **CallerRunsPolicy策略：**只要线程池未关闭，该策略直接在调用者线程中，运行当前被丢弃的任务。显然这样做不会真的丢弃任务，但是，任务提交线程的性能极有可能会急剧下降。
- **DiscardOledestPolicy策略：**该策略将丢弃最老的一个请求，也就是即将被执行的一个任务，并尝试再次提交当前任务。
- **DiscardPolicy策略：**该策略默默地丢弃无法处理的任务，不予任何处理。如果允许任务丢失，我觉得这可能是最好的一种方案了吧！

以上内置的策略均实现了`RejectedExecutionHandler`接口，若以上策略仍无法满足实际应用需要，完全可以自己扩展`RejectedExecutionHandler`接口。`RejectedExecutionHandler`的定义如下：

```
public interface RejectedExecutionHandler {  
    void rejectedExecution(Runnable r, ThreadPoolExecutor executor);  
}
```

其中r为请求执行的任务，executor为当前的线程池。

下面的代码简单地演示了自定义线程池和拒绝策略的使用：

```
01 public class RejectThreadPoolDemo {
02     public static class MyTask implements Runnable {
03         @Override
04         public void run() {
05             System.out.println(System.currentTimeMillis() + ":
06                 + Thread.currentThread().getId());
07             try {
08                 Thread.sleep(100);
09             } catch (InterruptedException e) {
10                 e.printStackTrace();
11             }
12         }
13     }
14
15     public static void main(String[] args) throws InterruptedE
16         MyTask task = new MyTask();
17         ExecutorService es = new ThreadPoolExecutor(5, 5,
18             0L, TimeUnit.MILLISECONDS,
19             new LinkedBlockingQueue<Runnable>(10),
20             Executors.defaultThreadFactory(),
21             new RejectedExecutionHandler(){
22                 @Override
23                 public void rejectedExecution(Runnable r,
```

```

24             ThreadPoolExecutor executor) {
25                 System.out.println(r.toString()+" is d
26             }
27         });
28         for (int i = 0; i < Integer.MAX_VALUE; i++) {
29             es.submit(task);
30             Thread.sleep(10);
31         }
32     }
33 }

```

上述代码的第17~27行自定义了一个线程池。该线程池有5个常驻线程，并且最大线程数量也是5个。这和固定大小的线程池是一样的。但是它却拥有一个只有10个容量的等待队列。因为使用无界队列很可能并不是最佳解决方案，如果任务量极大，很有可能会把内存撑爆。给出一个合理的队列大小，也是合乎常理的选择。同时，这里自定义了拒绝策略，我们不抛出异常，因为万一在任务提交端没有进行异常处理，则有可能使得整个系统都崩溃，这极有可能不是我们希望遇到的。但作为必要的信息记录，我们将任务丢弃的信息进行打印，当然，这只比内置的DiscardPolicy策略高级那么一点点。

由于在这个案例中，MyTask执行需要花费100毫秒，因此，必然会导致大量的任务被直接丢弃。执行上述代码，可能的部分输出如下：

```

1426597264669:Thread ID:11
1426597264679:Thread ID:12
java.util.concurrent.FutureTask@a57993 is discard
java.util.concurrent.FutureTask@1b84c92 is discard

```

可以看到，在执行几个任务后，拒绝策略就开始生效了。在实际应用中，我们可以将更详细的信息记录到日志中，来分析系统的负载和任务丢失的情况。

3.2.5 自定义线程创建： **ThreadFactory**

看了那么多有关线程池的介绍，不知道大家有没有思考过一个基本的问题：那就是线程池中的线程是从哪里来的呢？

之前我们介绍过，线程池的主要作用是为了线程复用，也就是避免了线程的频繁创建。但是，最开始的那些线程从何而来呢？答案就是ThreadFactory。

ThreadFactory是一个接口，它只有一个方法，用来创建线程：

```
Thread newThread(Runnable r);
```

当线程池需要新建线程时，就会调用这个方法。

自定义线程池可以帮助我们做不少事。比如，我们可以跟踪线程池究竟在何时创建了多少线程，也可以自定义线程的名称、组以及优先级等信息，甚至可以任性地将所有的线程设置为守护线程。总之，使用自定义线程池可以让我们更加自由地设置池子中所有线程的状态。下面的案例使用自定义的ThreadFactory，一方面记录了线程的创建，另一方面将所有的线程都设置为守护线程，这样，当主线程退出后，将会强制销毁线程池。

```
01 public static void main(String[] args) throws InterruptedException
02     MyTask task = new MyTask();
03     ExecutorService es = new ThreadPoolExecutor(5, 5,
04         0L, TimeUnit.MILLISECONDS,
05         new SynchronousQueue<Runnable>(),
06         new ThreadFactory(){
07             @Override
08             public Thread newThread(Runnable r) {
09                 Thread t= new Thread(r);
10                 t.setDaemon(true);
11                 System.out.println("create "+t);
12                 return t;
13             }
14         }
15     );
16     for (int i = 0; i < 5; i++) {
17         es.submit(task);
18     }
19     Thread.sleep(2000);
20 }
```

3.2.6 我的应用我做主：扩展线程池

虽然JDK已经帮我们实现了这个稳定的高性能线程池。但如果我们需要对这个线程池做一些扩展，比如，我们想监控每个任务执行的开始和结束时间，或者其他一些自定义的增强功能，这时候应该怎么办呢？

一个好消息是：ThreadPoolExecutor也是一个可以扩展的线程池。它提供了beforeExecute()、afterExecute()和terminated()三个接口对线程池进行控制。

以beforeExecute()、afterExecute()为例，在ThreadPoolExecutor.Worker.runTask()方法内部提供了这样的实现：

```
boolean ran = false;
beforeExecute(thread, task);                                //运行前
try {
    task.run();                                              //运行任务
    ran = true;
    afterExecute(task, null);                                //运行结束
    ++completedTasks;
} catch (RuntimeException ex) {
    if (!ran)
        afterExecute(task, ex);                            //运行结束
    throw ex;
}
```

ThreadPoolExecutor.Worker是ThreadPoolExecutor的内部类，它是一个实现了Runnable接口的类。ThreadPoolExecutor线程池中的工作线程也正是Worker实例。Worker.runTask()方法会被线程池以多线程模式异步调用，即Worker.runTask()会同时被多个线程访问。因此其beforeExecute()、afterExecute()接口也将同时多线程访问。

在默认的ThreadPoolExecutor实现中，提供了空的beforeExecute()和afterExecute()实现。在实际应用中，可以对其进行扩展来实现对线程池

运行状态的跟踪，输出一些有用的调试信息，以帮助系统故障诊断，这对于多线程程序错误排查是很有帮助的。下面演示了对线程池的扩展，在这个扩展中，我们将记录每一个任务的执行日志。

```
01 public class ExtThreadPool {
02     public static class MyTask implements Runnable {
03         public String name;
04
05         public MyTask(String name) {
06             this.name = name;
07         }
08
09         @Override
10         public void run() {
11             System.out.println("正在执行" + ":Thread ID:" + Thread
12                 + ",Task Name=" + name);
13             try {
14                 Thread.sleep(100);
15             } catch (InterruptedException e) {
16                 e.printStackTrace();
17             }
18         }
19     }
20
21     public static void main(String[] args) throws InterruptedException
22
23         ExecutorService es = new ThreadPoolExecutor(5, 5, 0L, Ti
```

```
24         new LinkedBlockingQueue<Runnable>()) {
25             @Override
26             protected void beforeExecute(Thread t, Runnable r)
27                 System.out.println("准备执行: " + ((MyTask) r).r
28             }
29
30             @Override
31             protected void afterExecute(Runnable r, Throwable
32                 System.out.println("执行完成: " + ((MyTask) r).r
33             }
34
35             @Override
36             protected void terminated() {
37                 System.out.println("线程池退出");
38             }
39
40     };
41     for (int i = 0; i < 5; i++) {
42         MyTask task = new MyTask("TASK-GEYM-" + i);
43         es.execute(task);
44         Thread.sleep(10);
45     }
46     es.shutdown();
47 }
48 }
```

上述代码在第23~40行，扩展了原有的线程池，实现了

`beforeExecute()`、`afterExecute()`和`terminated()`三个方法。这三个方法分别用于记录一个任务的开始、结束和整个线程池的退出。在第42~43行，向线程池提交5个任务，为了有更清晰的日志，我们为每个任务都取了一个不同的名字。第43行使用`execute()`方法提交任务，细心的读者一定发现，在之前代码中，我们都使用了`submit()`方法提交，有关两者的区别，我们将在“5.5节Future模式”中详细介绍。

在提交完成后，调用`shutdown()`方法关闭线程池。这是一个比较安全的方法，如果当前正有线程在执行，`shutdown()`方法并不会立即暴力地终止所有任务，它会等待所有任务执行完成后，再关闭线程池，但它并不会等待所有线程执行完成后再返回，因此，可以简单地理解成`shutdown()`只是发送了一个关闭信号而已。但在`shutdown()`方法执行后，这个线程池就不能再接受其他新的任务了。

执行上述代码，可以得到类似以下的输出：

```
准备执行：TASK-GEYM-0
正在执行：Thread ID:8,Task Name=TASK-GEYM-0
准备执行：TASK-GEYM-1
正在执行：Thread ID:9,Task Name=TASK-GEYM-1
准备执行：TASK-GEYM-2
正在执行：Thread ID:10,Task Name=TASK-GEYM-2
准备执行：TASK-GEYM-3
正在执行：Thread ID:11,Task Name=TASK-GEYM-3
准备执行：TASK-GEYM-4
正在执行：Thread ID:12,Task Name=TASK-GEYM-4
执行完成：TASK-GEYM-0
执行完成：TASK-GEYM-1
```

执行完成: TASK-GEYM-2

执行完成: TASK-GEYM-3

执行完成: TASK-GEYM-4

线程池退出

可以看到，所有任务的执行前、执行后的时间点以及任务的名字都已经可以捕获了。这对于应用程序的调试和诊断是非常有帮助的。

3.2.7 合理的选择：优化线程池线程数量

线程池的大小对系统的性能有一定的影响。过大或者过小的线程数量都无法发挥最优的系统性能，但是线程池大小的确定也不需要做得非常精确，因为只要避免极大和极小两种情况，线程池的大小对系统的性能并不会影响太大。一般来说，确定线程池的大小需要考虑CPU数量、内存大小等因素。在《Java Concurrency in Practice》一书中给出了一个估算线程池大小的经验公式：

Ncpu = CPU的数量

Ucpu = 目标CPU的使用率， $0 \leq Ucpu \leq 1$

W/C = 等待时间与计算时间的比率

为保持处理器达到期望的使用率，最优的池的大小等于：

Nthreads = Ncpu * Ucpu * (1 + W/C)

在Java中，可以通过：

```
Runtime.getRuntime().availableProcessors()
```

取得可用的CPU数量。

3.2.8 堆栈去哪里了：在线程池中寻找堆栈

大家一定还记得在上一章中，我们详解介绍了一些幽灵般的错误。我想，码农的痛苦也莫过于此了。多线程本身就是非常容易引起这类错误的。如果你使用了线程池，那么这种幽灵错误可能会变得更加常见。

下面来看一个简单的案例，首先，我们有一个Runnable接口，它用来计算两个数的商：

```
public class DivTask implements Runnable {
    int a,b;
    public DivTask(int a,int b){
        this.a=a;
        this.b=b;
    }
    @Override
    public void run() {
        double re=a/b;
        System.out.println(re);
    }
}
```

如果程序运行了这个任务，那么我们期望它可以打印出给定两个数的商。现在我们构造几个这样的任务，希望程序可以为我们计算一组给定数组的商：

```
public static void main(String[] args) throws InterruptedException {
    ThreadPoolExecutor pools=new ThreadPoolExecutor(0, Integer.MAX_VALUE,
        0L, TimeUnit.SECONDS,
        new SynchronousQueue<Runnable>());

    for(int i=0;i<5;i++){
        pools.submit(new DivTask(100,i));
    }
}
```

上述代码将DivTask提交到线程池，从这个for循环来看，我们应该会得到5个结果，分别是100除以给定的i后的商。但如果你真的运行程序，你得到的全部结果是：

```
33.0
50.0
100.0
25.0
```

你没有看错！只有4个输出。也就说是程序漏算了一组数据！但更不幸的是，程序没有任何日志，没有任何错误提示，就好像一切都正常一样。在这个简单的案例中，只要你稍有经验，你就能发现，作为除数的i取到了0，这个缺失的值很可能是由于除以0导致的。但在稍复杂的业务场景中，这种错误足可以让你几天萎靡不振。

因此，使用线程池虽然是件好事，但是还是得处处留意这些“坑”。线程池很有可能会“吃”掉程序抛出的异常，导致我们对程序的错误一无所知。

异常堆栈对于程序员的重要性就好像指南针对于茫茫大海上的船只。没有指南针，船只只能更艰难地寻找方向，没有异常堆栈，排查问题时，也只能像大海捞针那样，慢慢琢磨了。我的一个领导曾经说过：最鄙视那些出错不打印异常堆栈的行为！我相信，任何一个得益于异常堆栈而快速定位问题的程序员来说，一定对这句话深有体会。所以，这里我们将和大家讨论向线程池讨回异常堆栈的方法。

一种最简单的方法，就是放弃`submit()`，改用`execute()`。将上述的任务提交代码改成：

```
pools.execute(new DivTask(100,i));
```

或者你使用下面的方法改造你的`submit()`：

```
Future re=pools.submit(new DivTask(100,i));  
re.get();
```

上面两种方法都可以得到部分堆栈信息，如下所示：

```
Exception in thread "pool-1-thread-1" java.lang.ArithmeticExcepti  
    at geym.conc.ch3.trace.DivTask.run(DivTask.java:11)  
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoo  
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPo  
    at java.lang.Thread.run(Thread.java:745)
```

33.0

100.0

50.0

25.0

注意了，我这里说的是部分。这是因为从这两个异常堆栈中我们只能知道异常是在哪里抛出的（这里是DivTask的第11行）。但是我们还希望得到另外一个更重要的信息，那就是这个任务到底是在哪里提交的？而任务的具体提交位置已经被线程池完全淹没了。顺着堆栈，我们最多只能找到线程池中的调度流程，而这对于我们几乎是没有价值的。

既然这样，我们只能自己动手，丰衣足食啦！为了今后少加几天班，我们还是非常有必要将堆栈的信息彻底挖出来！扩展我们的ThreadPoolExecutor线程池，让它在调度任务之前，先保存一下提交任务线程的堆栈信息。如下所示：

```
01 public class TraceThreadPoolExecutor extends ThreadPoolExecuto
02     public TraceThreadPoolExecutor(int corePoolSize, int maxim
03         long keepAliveTime, TimeUnit unit, BlockingQueue<F
04         super(corePoolSize, maximumPoolSize, keepAliveTime, un
05     }
06
07     @Override
08     public void execute(Runnable task) {
09         super.execute(wrap(task, clientTrace(), Thread.current
10             .getName()));
11     }
12
13     @Override
```

```

14     public Future<?> submit(Runnable task) {
15         return super.submit(wrap(task, clientTrace(), Thread.c
16             .getName()));
17     }
18
19     private Exception clientTrace() {
20         return new Exception("Client stack trace");
21     }
22
23     private Runnable wrap(final Runnable task, final Exception
24         String clientThreadName) {
25         return new Runnable() {
26             @Override
27             public void run() {
28                 try {
29                     task.run();
30                 } catch (Exception e) {
31                     clientStack.printStackTrace();
32                     throw e;
33                 }
34             }
35         };
36     }
37 }

```

在第23行代码中，`wrap()`方法的第2个参数为一个异常，里面保存着提交任务的线程的堆栈信息。该方法将我们传入的`Runnable`任务进行一

层包装，使之能处理异常信息。当任务发生异常时，这个异常会被打印。

好了，现在可以使用我们的新成员（TraceThreadPoolExecutor）来尝试执行这段代码了：

```
14 public static void main(String[] args) {
15     ThreadPoolExecutor pools=new TraceThreadPoolExecutor(0, In
16         0L, TimeUnit.SECONDS,
17         new SynchronousQueue<Runnable>());
18
19     /**
20      * 错误堆栈中可以看到是在哪里提交的任务
21      */
22     for(int i=0;i<5;i++){
23         pools.execute(new DivTask(100,i));
24     }
25 }
```

执行上述代码，就可以得到以下信息：

```
java.lang.Exception: Client stack trace
    at geym.conc.ch3.trace.TraceThreadPoolExecutor.clientTrace(Tra
    at geym.conc.ch3.trace.TraceThreadPoolExecutor.execute(TraceTh
    at geym.conc.ch3.trace.TraceMain.main(TraceMain.java:23)
Exception in thread "pool-1-thread-1" java.lang.ArithmeticExcepti
    at geym.conc.ch3.trace.DivTask.run(DivTask.java:11)
    at geym.conc.ch3.trace.TraceThreadPoolExecutor$1.run(TraceThre
```

```
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
33.0
100.0
25.0
50.0
```

熟悉的异常又回来了！现在，我们不仅可以得到异常发生的Runnable实现内的信息，我们也知道了这个任务是在哪里提交的。如此丰富的信息，我相信可以帮助我们瞬间定位问题！

3.2.9 分而治之：Fork/Join框架

“分而治之”一直是一个非常有效地处理大量数据的方法。著名的MapReduce也是采取了分而治之的思想。简单来说，就是如果你要处理1000个数据，但是你并不具备处理1000个数据的能力，那么你可以只处理其中的10个，然后，分阶段处理100次，将100次的结果进行合成，那就是最终想要的对原始1000个数据的处理结果。

Fork一词的原始含义是吃饭用的叉子，也有分叉的意思。在Linux平台中，函数fork()用来创建子进程，使得系统进程可以多一个执行分支。在Java中也沿用了类似的命名方式。

而join()的含义在之前的章节中已经解释过，这里也是相同的意思，表示等待。也就是使用fork()后系统多了一个执行分支（线程），所以需要等待这个执行分支执行完毕，才有可能得到最终的结果，因此

join()就表示等待。

在实际使用中，如果毫无顾忌地使用fork()开启线程进行处理，那么很有可能导致系统开启过多的线程而严重影响性能。所以，在JDK中，给出了一个ForkJoinPool线程池，对于fork()方法并不急着开启线程，而是提交给ForkJoinPool线程池进行处理，以节省系统资源。使用Fork/Join进行数据处理时的总体结构如图3.8所示。

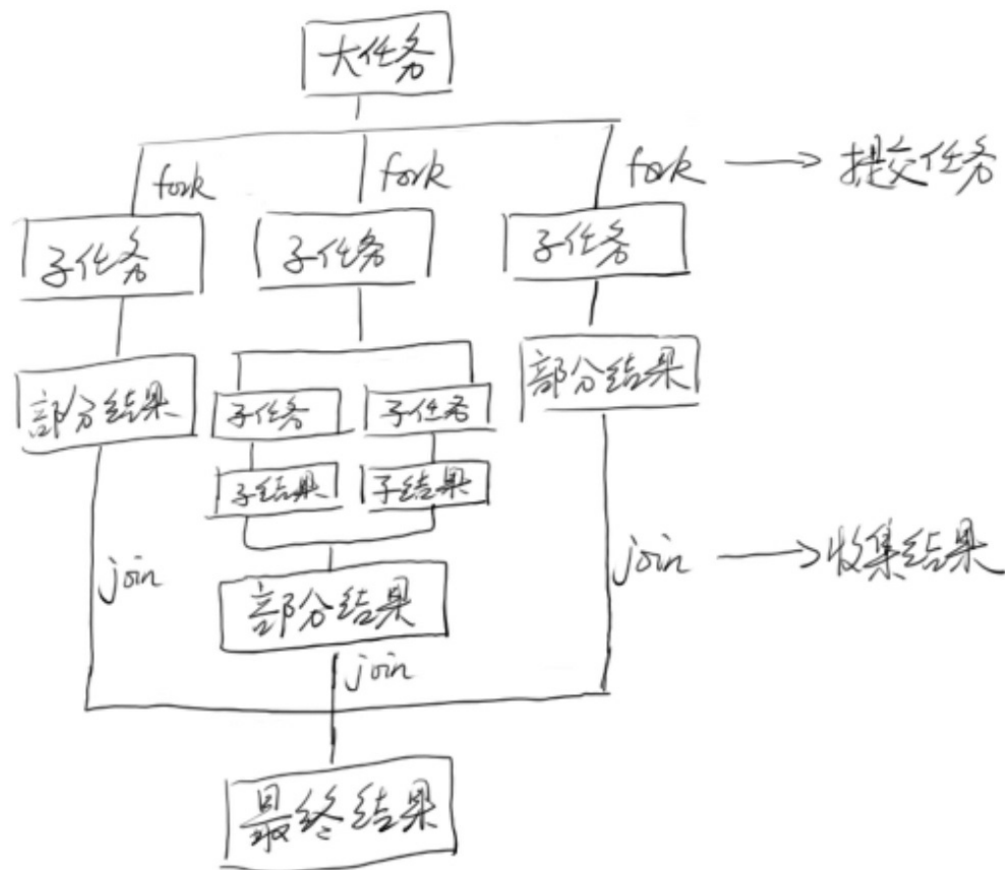


图3.8 Fork/Join执行逻辑

由于线程池的优化，提交的任务和线程数量并不是一对一的关系。在绝大多数情况下，一个物理线程实际上是需要处理多个逻辑任务的。因此，每个线程必然需要拥有一个任务队列。因此，在实际执行过程中，可能遇到这么一种情况：线程A已经把自己的任务都执行完成了，

而线程B还有一堆任务等着处理，此时，线程A就会“帮助”线程B，从线程B的任务队列中拿一个任务过来处理，尽可能地达到平衡。如图3.9所示，显示了这种互相帮助的精神。一个值得注意的地方是，当线程试图帮助别人时，总是从任务队列的底部开始拿数据，而线程试图执行自己的任务时，则是从相反的顶部开始拿。因此这种行为也十分有利于避免数据竞争。

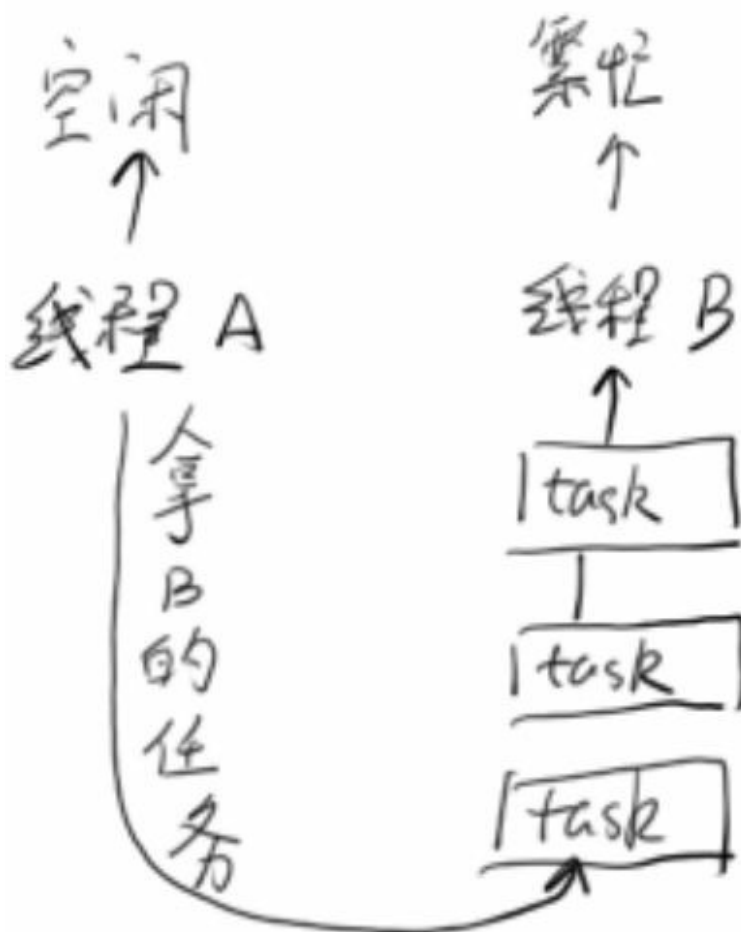


图3.9 互相帮助的线程

下面我们来看一下ForkJoinPool的一个重要的接口：

```
public <T> ForkJoinTask<T> submit(ForkJoinTask<T> task)
```

你可以向ForkJoinPool线程池提交一个ForkJoinTask任务。所谓ForkJoinTask任务就是支持fork()分解以及join()等待的任务。ForkJoinTask有两个重要的子类，RecursiveAction和RecursiveTask。它们分别表示没有返回值的任务和可以携带返回值的任务。图3.10显示了这两个类的作用和区别。

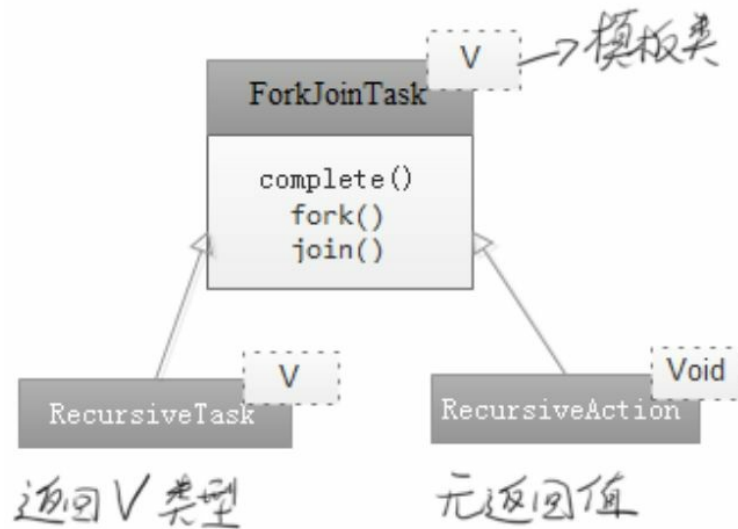


图3.10 RecursiveAction和RecursiveTask

下面我们简单地展示Fork/Join框架的使用，这里用来计算数列求和。

```
01 public class CountTask extends RecursiveTask<Long>{
02     private static final int THRESHOLD = 10000;
03     private long start;
04     private long end;
05
06     public CountTask(long start,long end){
07         this.start=start;
08         this.end=end;
```

```
09     }
10
11     public Long compute(){
12         long sum=0;
13         boolean canCompute = (end-start)<THRESHOLD;
14         if(canCompute){
15             for(long i=start;i<=end;i++){
16                 sum +=i;
17             }
18         }else{
19             //分成100个小任务
20             long step=(start+end)/100;
21             ArrayList<CountTask> subTasks=new ArrayList<CountTask>();
22             long pos=start;
23             for(int i=0;i<100;i++){
24                 long lastOne=pos+step;
25                 if(lastOne>end)lastOne=end;
26                 CountTask subTask=new CountTask(pos,lastOne);
27                 pos+=step+1;
28                 subTasks.add(subTask);
29                 subTask.fork();
30             }
31             for(CountTask t:subTasks){
32                 sum+=t.join();
33             }
34         }
35         return sum;
```



```

36     }
37
38     public static void main(String[]args){
39         ForkJoinPool forkJoinPool = new ForkJoinPool();
40         CountTask task = new CountTask(0,200000L);
41         ForkJoinTask<Long> result = forkJoinPool.submit(task)
42         try{
43             long res = result.get();
44             System.out.println("sum="+res);
45         }catch(InterruptedException e){
46             e.printStackTrace();
47         }catch(ExecutionException e){
48             e.printStackTrace();
49         }
50     }
51 }

```

由于计算数列的和必然是需要函数返回值的，因此选择RecursiveTask作为任务的模型。上述代码第39行，建立ForkJoinPool线程池。在第40行，构造一个计算1到200000求和的任务。在第41行将任务提交给线程池，线程池会返回一个携带结果的任务，通过get()方法可以得到最终结果（第43行）。如果在执行get()方法时，任务没有结束，那么主线程就会在get()方法时等待。

下面来看一下CountTask的实现。首先CountTask继承自RecursiveTask，可以携带返回值，这里的返回值类型设置为long。第2行定义的THRESHOLD设置了任务分解的规模，也就是如果要求和的

总数大于THRESHOLD个，那么任务就需要再次分解，否则就可以直接执行。这个判断逻辑在第14行有体现。如果任务可以直接执行，那么直接进行求和，返回结果。否则，就对任务再次分解。每次分解时，简单地将原有任务划分成100个等规模的小任务，并使用fork()提交子任务。之后，等待所有的子任务结束，并将结果再次求和（第31~33行）。

在使用ForkJoin时需要注意，如果任务的划分层次很深，一直得不到返回，那么可能出现两种情况：第一，系统内的线程数量越积越多，导致性能严重下降。第二，函数的调用层次变得很深，最终导致栈溢出。不同版本的JDK内部实现机制可能有差异，从而导致其表现不同。

下面的StackOverflowError异常就是加深本例的调用层次，在JDK 8上得到的错误。

```
java.util.concurrent.ExecutionException: java.lang.StackOverflowError
    at java.util.concurrent.ForkJoinTask.get(ForkJoinTask.java:10
    at geym.conc.ch3.fork.CountTask.main(CountTask.java:51)
Caused by: java.lang.StackOverflowError
```

此外，ForkJoin线程池使用一个无锁的栈来管理空闲线程。如果一个工作线程暂时取不到可用的任务，则可能会被挂起，挂起的线程将会被压入由线程池维护的栈中。待将来有任务可用时，再从栈中唤醒这些线程。

3.3 不要重复发明轮子：JDK的并发容器

除了提供诸如同步控制，线程池等基本工具外，为了提高开发人员的效率，JDK还为大家准备了一大批好用的容器类，可以大大减少开发工作量。大家应该都听说过一种说法，所谓程序就是“算法+数据结构”，这些容器类就是为大家准备好的线程数据结构。你可以在里面找到链表、HashMap、队列等。当然，它们都是线程安全的。

在这里，我也打算花一些篇幅为大家介绍一下这些工具类。这些容器类的封装都是非常完善并且“平易近人”的，也就是说只要你有那么一点点的编程经验，就可以非常容易地使用这些容器。因此，我可能会花更多的时间来分析这些工具的具体实现，希望起到抛砖引玉的作用。

3.3.1 超好用的工具类：并发集合简介

JDK提供的这些容器大部分在`java.util.concurrent`包中。我先提纲挈领地介绍一下它们，初次露脸，大家只需要知道它们的作用即可。有关具体的实现和注意事项，在后面我会慢慢道来。

- **ConcurrentHashMap**：这是一个高效的并发HashMap。你可以理解为一个线程安全的HashMap。
- **CopyOnWriteArrayList**：这是一个List，从名字看就是和ArrayList

是一族的。在读多写少的场合，这个List的性能非常好，远远好于Vector。

- **ConcurrentLinkedQueue**：高效的并发队列，使用链表实现。可以看做一个线程安全的LinkedList。
- **BlockingQueue**：这是一个接口，JDK内部通过链表、数组等方式实现了这个接口。表示阻塞队列，非常适合用于作为数据共享的通道。
- **ConcurrentSkipListMap**：跳表的实现。这是一个Map，使用跳表的数据结构进行快速查找。

除了以上并发包中的专有数据结构外，java.util下的Vector是线程安全的（虽然性能和上述专用工具没得比），另外Collections工具类可以帮助我们将任意集合包装成线程安全的集合。

3.3.2 线程安全的HashMap

在之前的章节中，已经给大家展示了在多线程环境中使用HashMap所带来的问题。那如果需要一个线程安全的HashMap应该怎么做呢？一种可行的方法是使用Collections.synchronizedMap()方法包装我们的HashMap。如下代码，产生的HashMap就是线程安全的：

```
public static Map m=Collections.synchronizedMap(new HashMap());
```

Collections.synchronizedMap()会生成一个名为SynchronizedMap的Map。它使用委托，将自己所有Map相关的功能交给传入的HashMap实现，而自己则主要负责保证线程安全。

具体参考下面的实现，首先SynchronizedMap内包装了一个Map。

```
private static class SynchronizedMap<K,V>
    implements Map<K,V>, Serializable {
    private static final long serialVersionUID = 197819847965

    private final Map<K,V> m;        // Backing Map
    final Object      mutex;        // Object on which to syn
```

通过mutex实现对这个m的互斥操作。比如，对于Map.get()方法，它的实现如下：

```
public V get(Object key) {
    synchronized (mutex) {return m.get(key);}
}
```

而其他所有相关的Map操作都会使用这个mutex进行同步。从而实现线程安全。

这个包装的Map可以满足线程安全的要求。但是，它在多线程环境中的性能表现并不算太好。无论是对Map的读取或者写入，都需要获得mutex的锁，这会导致所有对Map的操作全部进入等待状态，直到mutex锁可用。如果并发级别不高，一般也够用。但是，在高并发环境中，我们也有必要寻求新的解决方案。

一个更加专业的并发HashMap是ConcurrentHashMap。它位于java.util.concurrent包内。它专门为并发进行了性能优化，因此，更加适合多线程的场合。

有关ConcurrentHashMap的具体实现细节，大家可以参考“第4章锁的优化及注意事项”一章。我们将在那里给出更加详细的实现说明。

3.3.3 有关List的线程安全

队列、链表之类的数据结构也是极其常用的，几乎所有的应用程序都会与之相关。在Java中，ArrayList和Vector都是使用数组作为其内部实现。两者最大的不同在于Vector是线程安全的，而ArrayList不是。此外，LinkedList使用链表的数据结构实现了List。但是很不幸，LinkedList并不是线程安全的，不过参考前面对HashMap的包装，在这里我们也可以使用Collections.synchronizedList()方法来包装任意List，如下所示：

```
public static List<String> l=Collections.synchronizedList(new Li
```

此时生成的List对象就是线程安全的。

3.3.4 高效读写的队列：深度剖析 ConcurrentLinkedQueue

队列Queue也是常用的数据结构之一。在JDK中提供了一个ConcurrentLinkedQueue类用来实现高并发的队列。从名字可以看到，这个队列使用链表作为其数据结构。有关ConcurrentLinkedQueue的性能测试，大家可以自行尝试。这里限于篇幅就不再给出性能测试的代码。大家只要知道ConcurrentLinkedQueue应该算是在高并发环境中性能最好的队列就可以了。它之所以能有很好的性能，是因为其内部复杂的实现。

在这里，我更加愿意花一些篇幅来简单介绍一下ConcurrentLinkedQueue的具体实现细节。不过在深入ConcurrentLinkedQueue之前，我强烈建议大家先阅读一下第4章，补充一下有关无锁操作的一些知识。

作为一个链表，自然需要定义有关链表内的节点，在ConcurrentLinkedQueue中，定义的节点Node核心如下：

```
private static class Node<E> {  
    volatile E item;  
    volatile Node<E> next;  
}
```

其中item是用来表示目标元素的。比如，当列表中存放String时，item就是String类型。字段next表示当前Node的下一个元素，这样每个Node就能环环相扣，串在一起了。如图3.11所示，显示了ConcurrentLinkedQueue的基本结构。

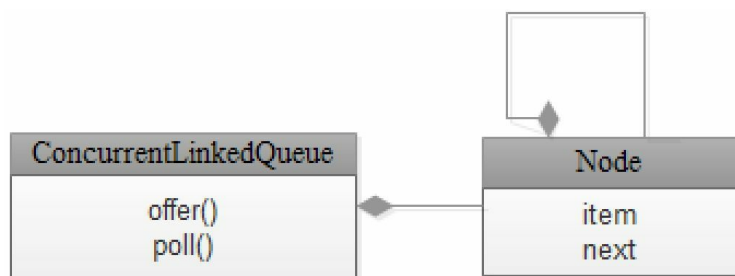


图3.11 ConcurrentLinkedQueue基本结构

对Node进行操作时，使用了CAS操作。

```
boolean casItem(E cmp, E val) {  
    return UNSAFE.compareAndSwapObject(this, itemOffset, cmp, val)  
}
```

```
void lazySetNext(Node<E> val) {
    UNSAFE.putOrderedObject(this, nextOffset, val);
}

boolean casNext(Node<E> cmp, Node<E> val) {
    return UNSAFE.compareAndSwapObject(this, nextOffset, cmp, val);
}
```

方法casItem()表示设置当前Node的item值。它需要两个参数，第一个参数为期望值，第二个参数为设置目标值。当当前值等于cmp期望值时，就会将目标设置为val。同样casItem()方法也是类似的，但是它是用来设置next字段，而不是item字段。

ConcurrentLinkedQueue内部有两个重要的字段，head和tail，分别表示链表的头部和尾部，它们都是Node类型。对于head来说，它永远不会为null，并且通过head以及succ()后继方法一定能完整地遍历整个链表。对于tail来说，它自然应该表示队列的末尾。

但ConcurrentLinkedQueue的内部实现非常复杂，它允许在运行时链表处于多个不同的状态。以tail为例，一般来说，我们期望tail总是为链表的末尾，但实际上，tail的更新并不是及时的，而是可能会产生拖延现象。如图3.12所示，显示了插入时，tail的更新情况，可以看到tail的更新会产生滞后，并且每次更新会跳跃两个元素。

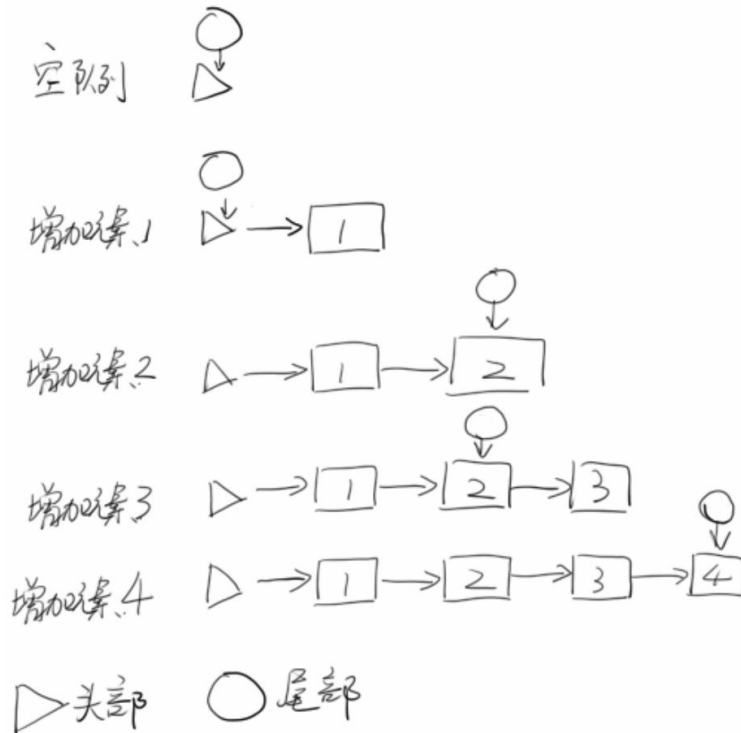


图3.12 插入节点时tail的更新

可以看到tail并不总是在更新。下面就是ConcurrentLinkedQueue中向队列中添加元素的offer()方法（本节中使用JDK 7u40的代码，不同版本的代码可能存在差异）：

```

01 public boolean offer(E e) {
02     checkNotNull(e);
03     final Node<E> newNode = new Node<E>(e);
04
05     for (Node<E> t = tail, p = t;;) {
06         Node<E> q = p.next;
07         if (q == null) {
08             // p 是最后一个节点
09             if (p.casNext(null, newNode)) {
10                 //每2次，更新一下tail

```

```

11         if (p != t)
12             casTail(t, newNode);
13             return true;
14     }
15     //CAS竞争失败，再次尝试
16 }
17 else if (p == q)
18     //遇到哨兵节点，从都head开始遍历。
19     //但如果tail被修改，则使用tail（因为可能被修改正确了）
20     p = (t != (t = tail)) ? t : head;
21 else
22     // 取下一个节点或者最后一个节点
23     p = (p != t && t != (t = tail)) ? t : q;
24 }
25 }

```

首先值得注意的是，这个方法没有任何锁操作。线程安全完全由CAS操作和队列的算法来保证。整个方法的核心是for循环，这个循环没有出口，直到尝试成功，这也符合CAS操作的流程。当第一次加入元素时，由于队列为空，因此p.next为null。程序进入第8行。并将p的next节点赋值为newNode，也就是将新的元素加入到队列中。此时p==t成立，因此不会执行第12行的代码更新tail末尾。如果casNext()成功，程序直接返回，如果失败，则再进行一次循环尝试，直到成功。因此，增加一个元素后，tail并不会被更新。

当程序试图增加第2个元素时，由于t还在head的位置上，因此p.next指向实际的第一个元素，因此第6行的q!=null，这表示q不是最后的节

点。由于往队列中增加元素需要最后一个节点的位置，因此，循环开始查找最后一个节点。于是，程序会进入第23行，获得最后一个节点。此时，`p`实际上是指向链表中的第一个元素，而它的`next`为`null`，故在第2个循环时，进入第8行。`p`更新自己的`next`，让它指向新加入的节点。如果成功，由于此时`p!=t`成功，则会更新`t`所在位置，将`t`移动到链表最后。

在第17行，处理了`p==q`的情况。这种情况是由于遇到了哨兵（`sentinel`）节点导致的。所谓哨兵节点，就是`next`指向自己的节点。这种节点在队列中的存在价值不大，主要表示要删除的节点，或者空节点。当遇到哨兵节点时，由于无法通过`next`取得后续的节点，因此很可能直接返回`head`，期望通过从链表头部开始遍历，进一步查找到链表末尾。但一旦发生在执行过程中，`tail`被其他线程修改的情况，则进行一次“打赌”，使用新的`tail`作为链表末尾（这样就避免了重新查找`tail`的开销）。

如果大家对Java不是特别熟悉，可能会对类似下面的代码产生疑惑（第20行）：

```
p = (t != (t = tail)) ? t : head;
```

这句代码虽然只有短短一行，但是包含的信息比较多。首先“`!=`”并不是原子操作，它是可以被中断的。也就是说，在执行“`!=`”是，程序会先取得`t`的值，再执行`t=tail`，并取得新的`t`的值。然后比较这两个值是否相等。在单线程时，`t!=t`这种语句显然不会成立。但是在并发环境中，有可能在获得左边的`t`值后，右边的`t`值被其他线程修改。这样，`t!=t`就可能成立。这里就是这种情况。如果在比较过程中，`tail`被其他线程修改，当它再次赋值给`t`时，就会导致等式左边的`t`和右边的`t`不同。如果两

个t不相同，表示tail在中途被其他线程篡改。这时，我们就可以用新的tail作为链表末尾，也就是这里等式右边的t。但如果tail没有被修改，则返回head，要求从头部开始，重新查找尾部。

作为简化问题，我们考察 $t!=t$ 的字节码（注意这里假设t为静态整形变量）：

```
11: getstatic    #10          // Field t:I
14: getstatic    #10          // Field t:I
17: if_icmpeq    24
```

可以看到，在字节码层面，t被先后取了两次，在多线程环境下，我们自然无法保证两次对t的取值会是相同的，如图3.13所示，显示了这种情况。

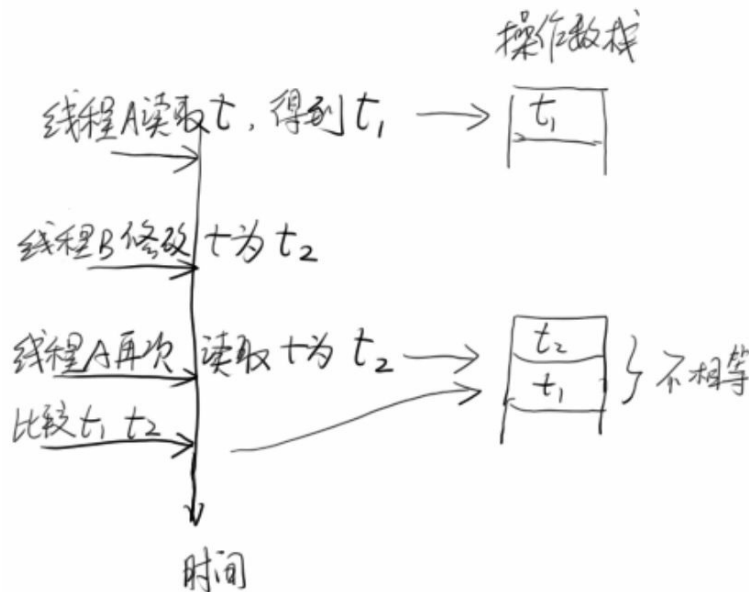


图3.13 $t!=t$ 成立的情况

下面我们来看一下哨兵节点是如何产生的：

```
ConcurrentLinkedQueue<String> q=new ConcurrentLinkedQueue<Strin
```

```
q.add("1");  
q.poll();
```

上述代码第3行，弹出队列内的元素。其执行过程如下：

```
01 public E poll() {  
02     restartFromHead:  
03     for (;;) {  
04         for (Node<E> h = head, p = h, q;;) {  
05             E item = p.item;  
06             if (item != null && p.casItem(item, null)) {  
07                 if (p != h)  
08                     updateHead(h, ((q = p.next) != null) ? q :  
09                     return item;  
10             }  
11             else if ((q = p.next) == null) {  
12                 updateHead(h, p);  
13                 return null;  
14             }  
15             else if (p == q)  
16                 continue restartFromHead;  
17             else  
18                 p = q;  
19         }  
20     }  
21 }
```

由于队列中只有一个元素，根据前文的描述，此时tail并没有更

新，而是指向和head相同的位置。而此时，head本身的item域为null，其next为列表第一个元素。故在第一个循环中，代码直接进入第18行，将p赋值为q，而q就是p.next，也是当前列表中的第一个元素。接着，在第2轮循环中，p.item显然不为null（为字符串1）。因此，代码应该可以顺利进入第7行（如果CAS操作成功）。进入第7行，也意味着p的item域被设置为null（因为这是弹出元素，自然需要删除）。同时，此时p和h是不相等的（因为p已经指向原有的第一个元素了）。故执行了第8行的updateHead()操作，其实现如下：

```
final void updateHead(Node<E> h, Node<E> p) {  
    if (h != p && casHead(h, p))  
        h.lazySetNext(h);  
}
```

可以看到，在updateHead中，就将p作为新的链表头部（通过casHead()实现），而原有的head就被设置为哨兵（通过lazySetNext()实现）。

这样一个哨兵节点就产生了，而由于此时原有的head头部和tail实际上是同一个元素。因此，再次offer()插入元素时，就会遇到这个tail，也就是哨兵。这就是offer()代码中，第17行的判断的意义。

通过这些说明，大家应该可以明显感觉到，不使用锁而单纯地使用CAS操作会要求应用层面保证线程安全，并处理一些可能存在的不一致问题，大大增加了程序设计和实现的难度。但是它带来的好处就是可以得到性能的飞速提升。因此，在有些场合也是值得的。

3.3.5 高效读取：不变模式下的 **CopyOnWriteArrayList**

在很多应用场景中，读操作可能会远远大于写操作。比如，有些系统级别的信息，往往只需要加载或者修改很少的次数，但是会被系统内所有模块频繁的访问。对于这种场景，我们最希望看到的就是读操作可以尽可能地快，而写即使慢一些也没有太大关系。

由于读操作根本不会修改原有的数据，因此对于每次读取都进行加锁其实是一种资源浪费。我们应该允许多个线程同时访问List的内部数据，毕竟读取操作是安全的。根据读写锁的思想，读锁和读锁之间确实也不冲突。但是，读操作会受到写操作的阻碍，当写发生时，读就必须等待，否则可能读到不一致的数据。同理，如果读操作正在进行，程序也不能进行写入。

为了将读取的性能发挥到极致，JDK中提供了CopyOnWriteArrayList类。对它来说，读取是完全不用加锁的，并且更好的消息是：写入也不会阻塞读取操作。只有写入和写入之间需要进行同步等待。这样一来，读操作的性能就会大幅度提升。那它是怎么做到的呢？

从这个类的名字我们可以看到，所谓CopyOnWrite就是在写入操作时，进行一次自我复制。换句话说，当这个List需要修改时，我并不修改原有的内容（这对于保证当前在读线程的数据一致性非常重要），而是对原有的数据进行一次复制，将修改的内容写入副本中。写完之后，再将修改完的副本替换原来的数据。这样就可以保证写操作不会影响读了。

下面的代码展示了有关读取的实现：

```
private volatile transient Object[] array;
public E get(int index) {
    return get(getArray(), index);
}
final Object[] getArray() {
    return array;
}
private E get(Object[] a, int index) {
    return (E) a[index];
}
```

需要注意的是：读取代码没有任何同步控制和锁操作，理由就是内部数组`array`不会发生修改，只会被另外一个`array`替换，因此可以保证数据安全。大家也可以参考“5.2不变模式”一节，相信可以有更深的认识。

和简单的读取相比，写入操作就有些麻烦了：

```
01 public boolean add(E e) {
02     final ReentrantLock lock = this.lock;
03     lock.lock();
04     try {
05         Object[] elements = getArray();
06         int len = elements.length;
07         Object[] newElements = Arrays.copyOf(elements, len + 1);
08         newElements[len] = e;
09         setArray(newElements);
```



```
10         return true;
11     } finally {
12         lock.unlock();
13     }
14 }
```

首先，写入操作使用锁，当然这个锁仅限于控制写-写的情况。其重点在于第7行代码，进行了内部元素的完整复制。因此，会生成一个新的数组newElements。然后，将新的元素加入newElements。接着，在第9行，使用新的数组替换老的数组，修改就完成了。整个过程不会影响读取，并且修改完后，读取线程可以立即“察觉”到这个修改（因为array变量是volatile类型）。

3.3.6 数据共享通道：BlockingQueue

前文中，我们已经提到了ConcurrentLinkedQueue作为高性能的队列。对于并发程序而言，高性能自然是一个我们需要追求的目标。但多线程的开发模式还会引入一个问题，那就是如何进行多个线程间的数据共享呢？比如，线程A希望给线程B发一个消息，用什么方式告知线程B是比较合理的呢？

一般来说，我们总是希望整个系统是松散耦合的。比如，你所在小区的物业希望可以得到一些业主的意见，设立了一个意见箱，如果对物业有任何要求和或者意见都可以投到意见箱里。这时，作为业主的你并不需要直接找到物业相关的领导表达你的意见。实际上，物业的工作人员也可能经常发生变动，直接找工作人员未必是一件方便的事情。而你投递到意见箱的意见总是会被物业的工作人员看到，不管是否发生了人

员的变动。这样，你就可以很容易地表达自己的诉求了。你既不需要直接和他们对话，又可以轻松提出自己的建议（这里假定我们物业公司的员工都是尽心尽责的好员工）。

将这个模式映射到我们程序中。就是说我们既希望线程A能够通知线程B，又希望线程A不知道线程B的存在。这样，如果将来进行重构或者升级，我们完全可以不修改线程A，而直接把线程B升级为线程C，保证系统的平滑过渡。而这中间的“意见箱”就可以使用BlockingQueue来实现。

与之前提到的ConcurrentLinkedQueue或者CopyOnWriteArrayList不同，BlockingQueue是一个接口，并非一个具体的实现。它的主要实现有下面一些，如图3.14所示。

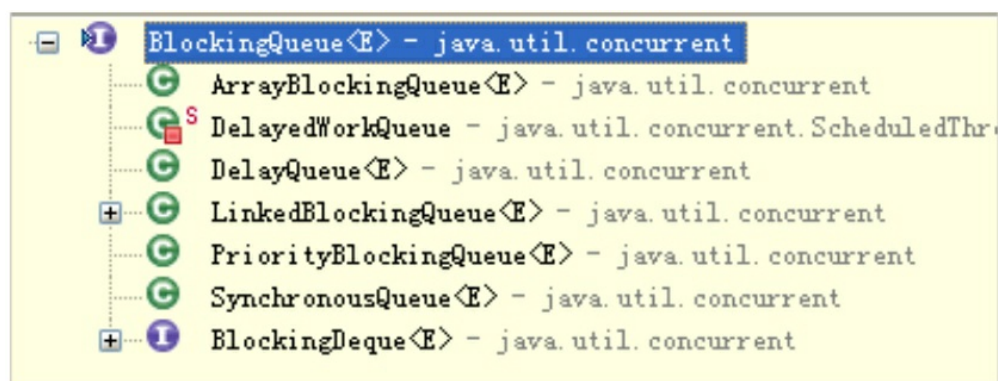


图3-14 BlockingQueue的主要实现

这里我们主要介绍ArrayBlockingQueue和LinkedBlockingQueue。从名字应该可以得知，ArrayBlockingQueue是基于数组实现的，而LinkedBlockingQueue基于链表。也正因为如此，ArrayBlockingQueue更适合做有界队列，因为队列中可容纳的最大元素需要在队列创建时指定（毕竟数组的动态扩展不太方便）。而LinkedBlockingQueue适合做无界队列，或者那些边界值非常大的队列，因为其内部元素可以动态增加，

它不会因为初值容量很大，而一口气吃掉你一大半的内存。

而BlockingQueue之所有适合作为数据共享的通道，其关键还在于Blocking上。Blocking是阻塞的意思，当服务线程（服务线程指不断获取队列中的消息，进行处理的线程）处理完成队列中所有的消息后，它如何知道下一条消息何时到来呢？

一种最傻瓜化的做法是让这个线程按照一定的时间间隔不停地循环和监控这个队列。这是可行的一种方案，但显然造成了不必要的资源浪费，而循环周期也难以确定。而BlockingQueue很好地解决了这个问题。它会让服务线程在队列为空时，进行等待，当有新的消息进入队列后，自动将线程唤醒，如图3.15所示。那它是如何实现的呢？我们以ArrayBlockingQueue为例，来一探究竟。

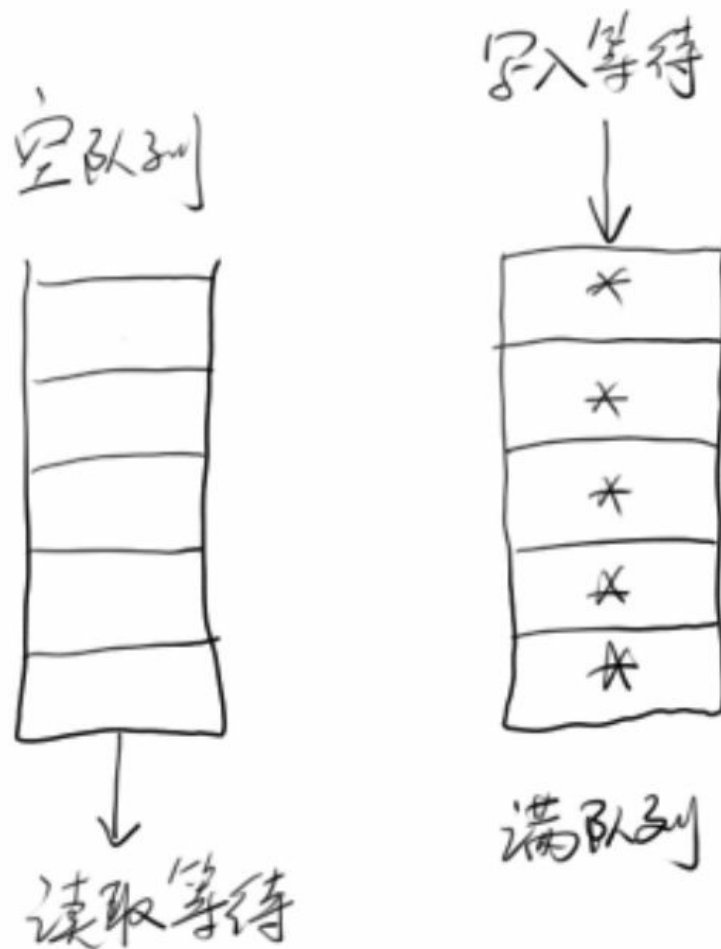


图3.15 BlockingQueue的工作模式

ArrayBlockingQueue的内部元素都放置在一个对象数组中：

```
final Object[] items;
```

向队列中压入元素可以使用offer()方法和put()方法。对于offer()方法，如果当前队列已经满了，它就会立即返回false。如果没有满，则执行正常的入队操作。所以，我们不讨论这个方法。现在，我们需要关注的是put()方法。put()方法也是将元素压入队列末尾。但如果队列满了，它会一直等待，直到队列中有空闲的位置。

从队列中弹出元素可以使用poll()方法和take()方法。它们都从队列

的头部获得一个元素。不同之处在于：如果队列为空poll()方法直接返回null，而take()方法会等待，直到队列内有可用元素。

因此，put()方法和take()方法才是体现Blocking的关键。为了做好等待和通知两件事，在ArrayBlockingQueue内部定义了以下一些字段：

```
final ReentrantLock lock;  
private final Condition notEmpty;  
private final Condition notFull;
```

当执行take()操作时，如果队列为空，则让当前线程等待在notEmpty上。新元素入队时，则进行一次notEmpty上的通知。

下面的代码显示了take()的过程：

```
01 public E take() throws InterruptedException {  
02     final ReentrantLock lock = this.lock;  
03     lock.lockInterruptibly();  
04     try {  
05         while (count == 0)  
06             notEmpty.await();  
07         return extract();  
08     } finally {  
09         lock.unlock();  
10     }  
11 }
```

第6行代码，就要求当前线程进行等待。当队列中有新元素时，线程会得到一个通知。下面是元素入队时的一段代码：

```
1 private void insert(E x) {
2     items[putIndex] = x;
3     putIndex = inc(putIndex);
4     ++count;
5     notEmpty.signal();
6 }
```

注意第5行代码，当新元素进入队列后，需要通知等待在notEmpty上的线程，让他们继续工作。

同理，对于put()操作也是一样的，当队列满时，需要让压入线程等待，如下面第7行。

```
01 public void put(E e) throws InterruptedException {
02     checkNotNull(e);
03     final ReentrantLock lock = this.lock;
04     lock.lockInterruptibly();
05     try {
06         while (count == items.length)
07             notFull.await();
08         insert(e);
09     } finally {
10         lock.unlock();
11     }
12 }
```

当有元素从队列中被挪走，队列中出现空位时，自然也需要通知等待入队的线程：

```
1 private E extract() {  
2     final Object[] items = this.items;  
3     E x = this.<E>cast(items[takeIndex]);  
4     items[takeIndex] = null;  
5     takeIndex = inc(takeIndex);  
6     --count;  
7     notFull.signal();  
8     return x;  
9 }
```

上述代码表示从队列中拿走一个元素。当有空闲位置时，在第7行，通知等待入队的线程。

BlockingQueue的使用非常普遍。在后续的“5.3生产者消费者”一节中，我们还会看到他们的身影。在那里，我们可以更清楚地看到如何使用BlockingQueue解耦生产者和消费者。

3.3.7 随机数据结构：跳表 (SkipList)

在JDK的并发包中，除了常用的哈希表外，还实现了一种有趣的数据结构——跳表。跳表是一种可以用来快速查找的数据结构，有点类似于平衡树。它们都可以对元素进行快速的查找。但一个重要的区别是：对平衡树的插入和删除往往很可能导致平衡树进行一次全局的调整。而对跳表的插入和删除只需要对整个数据结构的局部进行操作即可。这样带来的好处是：在高并发的情况下，你会需要一个全局锁来保证整个平

衡树的线程安全。而对于跳表，你只需要部分锁即可。这样，在高并发环境下，你就可以拥有更好的性能。而就查询的性能而言，跳表的时间复杂度也是 $O(\log n)$ 。所以在并发数据结构中，JDK使用跳表来实现一个Map。

跳表的另外一个特点是随机算法。跳表的本质是同时维护了多个链表，并且链表是分层的，如图3.16所示。

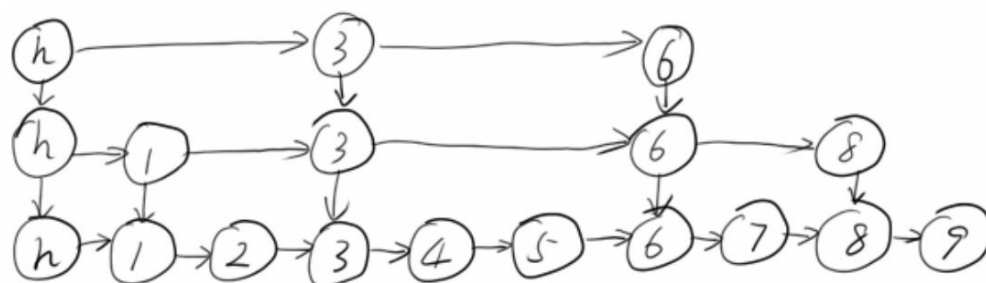


图3.16 跳表结构示意图

最低层的链表维护了跳表内所有的元素，每上面一层链表都是下面一层的子集，一个元素插入哪些层是完全随机的。因此，如果你运气不好的话，你可能会得到一个性能很糟糕的结构。但是在实际工作中，它的表现是非常好的。

跳表内的所有链表的元素都是排序的。查找时，可以从顶级链表开始找。一旦发现被查找的元素大于当前链表中的取值，就会转入下一层链表继续找。这也就是说在查找过程中，搜索是跳跃式的，如图3.17所示，在跳表中查找元素7。查找从顶层的头部索引节点开始。由于顶层的元素最少，因此，可以快速跳跃那些小于7的元素。很快，查找过程就能到元素6。由于在第2层，元素8大于7，故肯定无法在第2层找到元素7，故直接进入底层（包含所有元素）开始查找，并且很快就可以根据元素6搜索到元素7。整个过程，要比一般链表从元素1开始逐个搜索快很多。

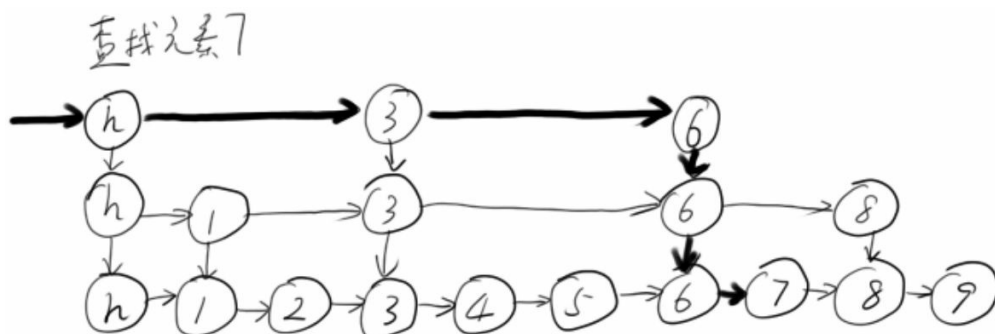


图3.17 跳表的查找过程

因此，很显然，跳表是一种使用空间换时间的算法。

使用跳表实现Map和使用哈希算法实现Map的另外一个不同之处是：哈希并不会保存元素的顺序，而跳表内所有的元素都是排序的。因此在对跳表进行遍历时，你会得到一个有序的结果。所以，如果你的应用需要有序性，那么跳表就是你不二的选择。

实现这一数据结构的类是ConcurrentSkipListMap。下面展示了跳表的简单使用：

```

Map<Integer, Integer> map=new ConcurrentSkipListMap<Integer, Integer>();
for(int i=0;i<30;i++){
    map.put(i,i);
}
for(Map.Entry<Integer, Integer> entry:map.entrySet()){
    System.out.println(entry.getKey());
}

```

和HashMap不同，对跳表的遍历输出是有序的。

跳表的内部实现有几个关键的数据结构组成。首先是Node，一个Node就是表示一个节点，里面含有两个重要的元素key和value（就是

Map的key和value)。每个Node还会指向下一个Node，因此还有一个元素next。

```
static final class Node<K,V> {  
    final K key;  
    volatile Object value;  
    volatile Node<K,V> next;
```

对Node的所有操作，使用的CAS方法：

```
boolean casValue(Object cmp, Object val) {  
    return UNSAFE.compareAndSwapObject(this, valueOffset, cmp, val);  
}  
  
boolean casNext(Node<K,V> cmp, Node<K,V> val) {  
    return UNSAFE.compareAndSwapObject(this, nextOffset, cmp, val);  
}
```

方法casValue()用来设置value的值，相对的casNext()用来设置next的字段。

另外一个重要的数据结构是Index。顾名思义，这个表示索引。它内部包装了Node，同时增加了向下的引用和向右的引用。

```
static class Index<K,V> {  
    final Node<K,V> node;  
    final Index<K,V> down;  
    volatile Index<K,V> right;
```

整个跳表就是根据Index进行全网的组织的。

此外，对于每一层的表头，还需要记录当前处于哪一层。为此，还需要一个称为HeadIndex的数据结构，表示链表头部的第一个Index。它继承自Index。

```
static final class HeadIndex<K,V> extends Index<K,V> {  
    final int level;  
    HeadIndex(Node<K,V> node, Index<K,V> down, Index<K,V> right)  
        super(node, down, right);  
    this.level = level;  
}  
}
```

这样核心的内部元素就介绍完成了。对于跳表的所有操作，就是组织好这些Index之间的连接关系。

3.4 参考资料

- 这篇博客讲解了ScheduledThreadPoolExecutor的使用注意事项
 - <http://segmentfault.com/a/1190000000371905>
- 这里讲解了几个有关线程池的使用技巧
 - <http://it.deepinmind.com/java/2014/11/26/executorservice-10-tips-and-tricks.html>
- 有关Fork/Join的简单实现原理
 - <http://www.infoq.com/cn/articles/fork-join-introduction>
- 有关ConcurrentLinkedQueue的实现具体分析（其使用的JDK版本与本书不同）
 - <http://my.oschina.net/xianggao/blog/389332>
 - <http://www.ibm.com/developerworks/cn/java/j-lo-concurrent/>
- 有关ConcurrentSkipListMap的运作原理（示例图示很好）
 - <http://www.liuhaihua.cn/archives/40657.html>

第4章 锁的优化及注意事项

“锁”是最常用的同步方法之一。在高并发的环境下，激烈的锁竞争会导致程序的性能下降。所以我们自然有必要讨论一些有关“锁”的性能问题以及相关一些注意事项。比如：避免死锁、减小锁粒度、锁分离等。

在多核时代，使用多线程可以明显地提高系统的性能。但事实上，使用多线程的方式会额外增加系统的开销。

对于单任务或者单线程的应用而言，其主要资源消耗都花在任务本身。它既不需要维护并行数据结构间的一致性状态，也不需要为线程的切换和调度花费时间。但对于多线程应用来说，系统除了处理功能需求外，还需要额外维护多线程环境的特有信息，如线程本身的元数据、线程的调度、线程上下文的切换等。

事实上，在单核CPU上，采用并行算法的效率一般要低于原始的串行算法的，其根本原因也在于此。因此，并行计算之所以能提高系统的性能，并不是因为它“少干活”了，而是因为并行计算可以更合理地进行任务调度，充分利用各个CPU资源。因此，合理的并发，才能将多核CPU的性能发挥到极致。

4.1 有助于提高“锁”性能的几点建议

“锁”的竞争必然会导致程序的整体性能下降。为了将这种副作用降到最低，我这里提出一些关于使用锁的建议，希望可以帮助大家写出性能更为优越的程序。

4.1.1 减小锁持有时间

对于使用锁进行并发控制的应用程序而言，在锁竞争过程中，单个线程对锁的持有时间与系统性能有着直接的关系。如果线程持有锁的时间很长，那么相对地，锁的竞争程度也就越激烈。可以想象一下，如果要求100个人各自填写自己的身份信息，但是只给他们一支笔。那么如果每个人拿着笔的时间都很长，总体所花的时间就会很长。如果真的只能有一支笔共享给100个人用，那么最好就让每个人花尽量少的时间持笔，务必做到想好了再拿笔写，千万不可拿着笔才去思考这表格应该怎么填。程序开发也是类似的，应该尽可能地减少对某个锁的占有时间，以减少线程间互斥的可能。以下面的代码段为例：

```
public synchronized void syncMethod(){
    othercode1();
    mutexMethod();
    othercode2();
}
```

syncMethod()方法中，假设只有mutextMethod()方法是有同步需要的，而othercode1()和othercode2()并不需要做同步控制。如果othercode1()和othercode2()分别是重量级的方法，则会花费较长的CPU时间。此时，如果在并发量较大，使用这种对整个方法做同步的方案，会导致等待线程大量增加。因为一个线程，在进入该方法时获得内部锁，只有在所有任务都执行完后，才会释放锁。

一个较为优化的解决方案是，只在必要时进行同步，这样就能明显减少线程持有锁的时间，提高系统的吞吐量。

```
public void syncMethod2(){
    othercode1();
    synchronized(this){
        mutextMethod();
    }
    othercode2();
}
```

在改进的代码中，只针对mutextMethod()方法做了同步，锁占用的时间相对较短，因此能有更高的并行度。这种技术手段在JDK的源码包中也可以很容易地找到，比如处理正则表达式的Pattern类：

```
public Matcher matcher(CharSequence input) {
    if (!compiled) {
        synchronized(this) {
            if (!compiled)
                compile();
        }
    }
}
```

```
}  
    Matcher m = new Matcher(this, input);  
    return m;  
}
```

`matcher()`方法有条件地进行锁申请，只有在表达式未编译时，进行局部的加锁。这种处理方式大大提高了`matcher()`方法的执行效率和可靠性。

注意：减少锁的持有时间有助于降低锁冲突的可能性，进而提升系统的并发能力。

4.1.2 减小锁粒度

减小锁粒度也是一种削弱多线程锁竞争的有效手段。这种技术典型的使用场景就是`ConcurrentHashMap`类的实现。大家应该还记得这个类吧！在“3.3 JDK的并发容器”一节中，我向大家介绍了这个高性能的`HashMap`。但是当时我们并没有说明它的实现原理。这里，让我们更加细致地看一下这个类。

对于`HashMap`来说，最重要的两个方法就是`get()`和`put()`。一种最自然的想法就是对整个`HashMap`加锁，必然可以得到一个线程安全的对象。但是这样做，我们就认为加锁粒度太大。对于`ConcurrentHashMap`，它内部进一步细分了若干个小的`HashMap`，称之为段（SEGMENT）。默认情况下，一个`ConcurrentHashMap`被进一步细分为16个段。

如果需要在ConcurrentHashMap中增加一个新的表项，并不是将整个HashMap加锁，而是首先根据hashCode得到该表项应该被存放到哪个段中，然后对该段加锁，并完成put()操作。在多线程环境中，如果多个线程同时进行put()操作，只要被加入的表项不存放在同一个段中，则线程间便可以做到真正的并行。

由于默认有16个段，因此，如果够幸运的话，ConcurrentHashMap可以同时接受16个线程同时插入（如果都插入不同的段中），从而大大提供其吞吐量。下面代码显示了put()操作的过程。在第5~6行，根据key，获得对应的段的序号。接着在第9行，得到段，然后将数据插入给定的段中。

```
01 public V put(K key, V value) {
02     Segment<K,V> s;
03     if (value == null)
04         throw new NullPointerException();
05     int hash = hash(key);
06     int j = (hash >>> segmentShift) & segmentMask;
07     if ((s = (Segment<K,V>)UNSAFE.getObject                //
08         (segments, (j << SSHIFT) + SBASE)) == null)      //
09         s = ensureSegment(j);
10     return s.put(key, hash, value, false);
11 }
```

但是，减少锁粒度会引入一个新的问题，即：当系统需要取得全局锁时，其消耗的资源会比较多。仍然以ConcurrentHashMap类为例，虽然其put()方法很好地分离了锁，但是当试图访问ConcurrentHashMap全局信息时，就会需要同时取得所有段的锁方能顺利实施。比如

ConcurrentHashMap的size()方法，它将返回ConcurrentHashMap的有效表项的数量，即ConcurrentHashMap的全部有效表项之和。要获取这个信息需要取得所有子段的锁，因此，其size()方法的部分代码如下：

```
sum = 0;
for (int i = 0; i < segments.length; ++i)           //对所有段
    segments[i].lock();
for (int i = 0; i < segments.length; ++i)           //统计总数
    sum += segments[i].count;
for (int i = 0; i < segments.length; ++i)           //释放所有段锁
    segments[i].unlock();
```

可以看到在计算总数时，先要获得所有段的锁，然后再求和。但是，ConcurrentHashMap的size()方法并不总是这样执行，事实上，size()方法会先使用无锁的方式求和，如果失败才会尝试这种加锁的方法。但不管这么说，在高并发场合ConcurrentHashMap的size()的性能依然要差于同步的HashMap。

因此，只有在类似于size()获取全局信息的方法调用并不频繁时，这种减小锁粒度的方法才能真正意义上提高系统吞吐量。

注意：所谓减少锁粒度，就是指缩小锁定对象的范围，从而减少锁冲突的可能性，进而提高系统的并发能力。

4.1.3 读写分离锁来替换独占锁

在之前我们已经提过，使用读写锁ReadWriteLock可以提高系统的

性能。使用读写分离锁来替代独占锁是减小锁粒度的一种特殊情况。如果说上节中提到的减少锁粒度是通过分割数据结构实现的，那么，读写锁则是对系统功能点的分割。

在读多写少的场合，读写锁对系统性能是很有好处的。因为如果系统在读写数据时均只使用独占锁，那么读操作和写操作间、读操作和读操作间、写操作和写操作间均不能做到真正的并发，并且需要相互等待。而读操作本身不会影响数据的完整性和一致性。因此，理论上讲，在大部分情况下，应该可以允许多线程同时读，读写锁正是实现了这种功能。由于我们在第3章中已经介绍了读写锁，因此这里就不再重复了。

注意：在读多写少的场合，使用读写锁可以有效提升系统的并发能力。

4.1.4 锁分离

如果将读写锁的思想做进一步的延伸，就是锁分离。读写锁根据读写操作功能上的不同，进行了有效的锁分离。依据应用程序的功能特点，使用类似的分离思想，也可以对独占锁进行分离。一个典型的案例就是`java.util.concurrent.LinkedBlockingQueue`的实现（如果大家印象深刻，我们在之前已经讨论了它的近亲`ArrayBlockingQueue`的内部实现）。

在`LinkedBlockingQueue`的实现中，`take()`函数和`put()`函数分别实现了从队列中取得数据和往队列中增加数据的功能。虽然两个函数都对当前队列进行了修改操作，但由于`LinkedBlockingQueue`是基于链表的，因

此，两个操作分别作用于队列的前端和尾端，从理论上说，两者并不冲突。

如果使用独占锁，则要求在两个操作进行时获取当前队列的独占锁，那么take()和put()操作就不可能真正的并发，在运行时，它们会彼此等待对方释放锁资源。在这种情况下，锁竞争会相对比较激烈，从而影响程序在高并发时的性能。

因此，在JDK的实现中，并没有采用这样的方式，取而代之的是两把不同的锁，分离了take()和put()操作。

```
/** Lock held by take, poll, etc */
private final ReentrantLock takeLock = new ReentrantLock();    //
/** Wait queue for waiting takes */
private final Condition notEmpty = takeLock.newCondition();
/** Lock held by put, offer, etc */
private final ReentrantLock putLock = new ReentrantLock();     //
/** Wait queue for waiting puts */
private final Condition notFull = putLock.newCondition();
```

以上代码片段，定义了takeLock和putLock，它们分别在take()操作和put()操作中使用。因此，take()函数和put()函数就此相互独立，它们之间不存在锁竞争关系，只需要在take()和take()间、put()和put()间分别对takeLock和putLock进行竞争。从而，削弱了锁竞争的可能性。

函数take()的实现如下，笔者在代码中给出了详细的注释，故不在正文中做进一步说明。

```
public E take() throws InterruptedException {
```

```

    E x;
    int c = -1;
    final AtomicInteger count = this.count;
    final ReentrantLock takeLock = this.takeLock;
    takeLock.lockInterruptibly();           //不能有两个
    try {
        try {
            while (count.get() == 0)       //如果当前没
                notEmpty.await();          //等待, put
        } catch (InterruptedException ie) {
            notEmpty.signal();             //通知其他未
            throw ie;
        }

        x = extract();                     //取得第一个
        c = count.getAndDecrement();       //数量减1, )
//函数同时访问count。注意: 变量c是
//count减1前的值
        if (c > 1)
            notEmpty.signal();            //通知其他ta
    } finally {
        takeLock.unlock();                //释放锁
    }
    if (c == capacity)
        signalNotFull();                  //通知put()
    return x;
}

```

函数put()的实现如下，

```
public void put(E e) throws InterruptedException {
    if (e == null) throw new NullPointerException();
    int c = -1;
    final ReentrantLock putLock = this.putLock;
    final AtomicInteger count = this.count;
    putLock.lockInterruptibly();                //不能有两个线程同时持有锁
    try {
        try {
            while (count.get() == capacity)    //如果队列已经满了
                notFull.await();                //等待
        } catch (InterruptedException ie) {
            notFull.signal();                    //通知未中断的线程
            throw ie;
        }
        insert(e);                               //插入数据
        c = count.getAndIncrement();            //更新总数，变为0
        if (c + 1 < capacity)
            notFull.signal();                    //有足够的空间
    } finally {
        putLock.unlock();                        //释放锁
    }
    if (c == 0)
        signalNotEmpty();                        //插入成功后，
}
```

通过takeLock和putLock两把锁，LinkedBlockingQueue实现了取数据和写数据的分离，使两者在真正意义上成为可并发的操作。

4.1.5 锁粗化

通常情况下，为了保证多线程间的有效并发，会要求每个线程持有锁的时间尽量短，即在使用完公共资源后，应该立即释放锁。只有这样，等待在这个锁上的其他线程才能尽早地获得资源执行任务。但是，凡事都有一个度，如果对同一个锁不停地进行请求、同步和释放，其本身也会消耗系统宝贵的资源，反而不利于性能的优化。

为此，虚拟机在遇到一连串连续地对同一锁不断进行请求和释放的操作时，便会把所有的锁操作整合成对锁的一次请求，从而减少对锁的请求同步次数，这个操作叫做锁的粗化。比如代码段：

```
public void demoMethod(){
    synchronized(lock){
        //do sth.
    }
    //做其他不需要的同步的工作，但能很快执行完毕
    synchronized(lock){
        //do sth.
    }
}
```

会被整合成如下形式：

```
public void demoMethod(){
    //整合成一次锁请求
    synchronized(lock){
        //do sth.
        //做其他不需要的同步的工作，但能很快执行完毕
    }
}
```

在开发过程中，大家也应该有意识地在合理的场合进行锁的粗化，尤其当在循环内请求锁时。以下是一个循环内请求锁的例子，在这种情况下，意味着每次循环都有申请锁和释放锁的操作。但在这种情况下，显然是没有必要的。

```
for(int i=0;i<CIRCLE;i++){
    synchronized(lock){

    }
}
```

所以，一种更加合理的做法应该是在外层只请求一次锁：

```
synchronized(lock){
for(int i=0;i<CIRCLE;i++){

}
}
```

注意：性能优化就是根据运行时的真实情况对各个资源点进行权衡

折中的过程。锁粗化的思想和减少锁持有时间是相反的，但在不同的场合，它们的效果并不相同。所以大家需要根据实际情况，进行权衡。

4.2 Java虚拟机对锁优化所做的努力

作为一款共用平台，JDK本身也为并发程序的性能绞尽脑汁。在JDK内部也想尽一切办法提供并发时的系统吞吐量。这里，我将向大家简单介绍几种JDK内部的“锁”优化策略。

4.2.1 锁偏向

锁偏向是一种针对加锁操作的优化手段。它的核心思想是：如果一个线程获得了锁，那么锁就进入偏向模式。当这个线程再次请求锁时，无须再做任何同步操作。这样就节省了大量有关锁申请的操作，从而提高了程序性能。因此，对于几乎没有锁竞争的场合，偏向锁有比较好的优化效果，因为连续多次极有可能是同一个线程请求相同的锁。而对于锁竞争比较激烈的场合，其效果不佳。因为在竞争激烈的场合，最有可能的情况是每次都是不同的线程来请求相同的锁。这样偏向模式会失效，因此还不如不启用偏向锁。使用Java虚拟机参数-XX:+UseBiasedLocking可以开启偏向锁。

4.2.2 轻量级锁

如果偏向锁失败，虚拟机并不会立即挂起线程。它还会使用一种称为轻量级锁的优化手段。轻量级锁的操作也很轻便，它只是简单地将对

象头部作为指针，指向持有锁的线程堆栈的内部，来判断一个线程是否持有对象锁。如果线程获得轻量级锁成功，则可以顺利进入临界区。如果轻量级锁加锁失败，则表示其他线程抢先争夺到了锁，那么当前线程的锁请求就会膨胀为重量级锁。

4.2.3 自旋锁

锁膨胀后，虚拟机为了避免线程真实地在操作系统层面挂起，虚拟机还会在做最后的努力——自旋锁。由于当前线程暂时无法获得锁，但是什么时候可以获得锁是一个未知数。也许在几个CPU时钟周期后，就可以得到锁。如果这样，简单粗暴地挂起线程可能是一种得不偿失的操作。因此，系统会进行一次赌注：它会假设在不久的将来，线程可以得到这把锁。因此，虚拟机会让当前线程做几个空循环（这也是自旋的含义），在经过若干次循环后，如果可以得到锁，那么就顺利进入临界区。如果还不能获得锁，才会真实地将线程在操作系统层面挂起。

4.2.4 锁消除

锁消除是一种更彻底的锁优化。Java虚拟机在JIT编译时，通过对运行上下文的扫描，去除不可能存在共享资源竞争的锁。通过锁消除，可以节省毫无意义的请求锁时间。

说到这里，细心的读者可能会产生疑问，如果不可能存在竞争，为什么程序员还要加上锁呢？这是因为在Java软件开发过程中，我们必然会使用一些JDK的内置API，比如StringBuffer、Vector等。你在使用这些类的时候，也许根本不会考虑这些对象到底内部是如何实现的。比

如，你很有可能在一个不可能存在并发竞争的场所使用Vector。而众所周知，Vector内部使用了synchronized请求锁。比如下面的代码：

```
public String[] createStrings(){
    Vector<String> v=new Vector<String>();
    for(int i=0;i<100;i++){
        v.add(Integer.toString(i));
    }
    return v.toArray(new String[]{});
}
```

注意上述代码中的Vector，由于变量v只在createStrings()函数中使用，因此，它只是一个单纯的局部变量。局部变量是在线程栈上分配的，属于线程私有的数据，因此不可能被其他线程访问。所以，在这种情况下，Vector内部所有加锁同步都是没有必要的。如果虚拟机检测到这种情况，就会将这些无用的锁操作去除。

锁消除涉及的一项关键技术为逃逸分析。所谓逃逸分析就是观察某一个变量是否会逃出某一个作用域。在本例中，变量v显然没有逃出createStrings()函数之外。以次为基础，虚拟机才可以大胆地将v内部的加锁操作去除。如果createStrings()返回的不是String数组，而是v本身，那么就认为变量v逃逸出了当前函数，也就是说v有可能被其他线程访问。如果是这样，虚拟机就不能消除v中的锁操作。

逃逸分析必须在-server模式下进行，可以使用-XX:+DoEscapeAnalysis参数打开逃逸分析。使用-XX:+EliminateLocks参数可以打开锁消除。

4.3 人手一支笔：ThreadLocal

除了控制资源的访问外，我们还可以通过增加资源来保证所有对象的线程安全。比如，让100个人填写个人信息表，如果只有一支笔，那么大家就得挨个填写，对于管理人员来说，必须保证大家不会去哄抢这仅存的一支笔，否则，谁也填不完。从另外一个角度出发，我们可以干脆就准备100支笔，人手一支，那么所有人都可以各自为营，很快就能完成表格的填写工作。

如果说锁是使用第一种思路，那么ThreadLocal就是使用第二种思路了。

4.3.1 ThreadLocal的简单使用

从ThreadLocal的名字上可以看到，这是一个线程的局部变量。也就是说，只有当前线程可以访问。既然是只有当前线程可以访问的数据，自然是线程安全的。

下面来看一个简单的示例：

```
01 private static final SimpleDateFormat sdf = new SimpleDateFormat
02 public static class ParseDate implements Runnable{
03     int i=0;
04     public ParseDate(int i){this.i=i;}
05     public void run() {
```

```

06         try {
07             Date t=sdf.parse("2015-03-29 19:29:"+i%60);
08             System.out.println(i+": "+t);
09         } catch (ParseException e) {
10             e.printStackTrace();
11         }
12     }
13 }
14 public static void main(String[] args) {
15     ExecutorService es=Executors.newFixedThreadPool(10);
16     for(int i=0;i<1000;i++){
17         es.execute(new ParseDate(i));
18     }
19 }

```

上述代码在多线程中使用SimpleDateFormat来解析字符串类型的日期。如果你执行上述代码，一般来说，你很可能得到一些异常（篇幅有限不再给出堆栈，只给出异常名称）：

```

Exception in thread "pool-1-thread-26" java.lang.NumberFormatExce
Exception in thread "pool-1-thread-17" java.lang.NumberFormatExce

```

出现这些问题的原因，是SimpleDateFormat.parse()方法并不是线程安全的。因此，在线程池中共享这个对象必然导致错误。

一种可行的方案是在sdf.parse()前后加锁，这也是我们一般的处理思路。这里我们不这么做，我们使用ThreadLocal为每一个线程都产生一个SimpleDateFormat对象实例：

```
01 static ThreadLocal<SimpleDateFormat> tl=new ThreadLocal<SimpleDateFormat>() {
02     public static class ParseDate implements Runnable{
03         int i=0;
04         public ParseDate(int i){this.i=i;}
05         public void run() {
06             try {
07                 if(tl.get()==null){
08                     tl.set(new SimpleDateFormat("yyyy-MM-dd HH:mm:ss"));
09                 }
10                 Date t=tl.get().parse("2015-03-29 19:29:"+i%60);
11                 System.out.println(i+": "+t);
12             } catch (ParseException e) {
13                 e.printStackTrace();
14             }
15         }
16     }
}
```

上述代码第7~9行，如果当前线程不持有SimpleDateFormat对象实例。那么就新建一个并把它设置到当前线程中，如果已经持有，则直接使用。

从这里也可以看到，为每一个线程人手分配一个对象的工作并不是由ThreadLocal来完成的，而是需要在应用层面保证的。如果在应用上为每一个线程分配了相同的对象实例，那么ThreadLocal也不能保证线程安全。这点也需要大家注意。

注意：为每一个线程分配不同的对象，需要在应用层面保证。

ThreadLocal只是起到了简单的容器作用。

4.3.2 ThreadLocal的实现原理

那ThreadLocal又是如何保证这些对象只被当前线程所访问呢？下面让我们一起深入ThreadLocal的内部实现。

我们需要关注的，自然是ThreadLocal的set()方法和get()方法。从set()方法先说起：

```
public void set(T value) {  
    Thread t = Thread.currentThread();  
    ThreadLocalMap map = getMap(t);  
    if (map != null)  
        map.set(this, value);  
    else  
        createMap(t, value);  
}
```

在set时，首先获得当前线程对象，然后通过getMap()拿到线程的ThreadLocalMap，并将值设入ThreadLocalMap中。而ThreadLocalMap可以理解为一个Map（虽然不是，但是你可以把它简单地理解成HashMap），但是它是定义在Thread内部的成员。注意下面的定义是从Thread类中摘出来的：

```
ThreadLocal.ThreadLocalMap threadLocals = null;
```

而设置到ThreadLocal中的数据，也正是写入了threadLocals这个Map。其中，key为ThreadLocal当前对象，value就是我们需要的值。而threadLocals本身就保存了当前自己所在线程的所有“局部变量”，也就是

一个ThreadLocal变量的集合。

在进行get()操作时，自然就是将这个Map中的数据拿出来：

```
public T get() {
    Thread t = Thread.currentThread();
    ThreadLocalMap map = getMap(t);
    if (map != null) {
        ThreadLocalMap.Entry e = map.getEntry(this);
        if (e != null)
            return (T)e.value;
    }
    return setInitialValue();
}
```

首先，get()方法也是先取得当前线程的ThreadLocalMap对象。然后，通过将自己作为key取得内部的实际数据。

在了解了ThreadLocal的内部实现后，我们自然会引出一个问题。那就是这些变量是维护在Thread类内部的（ThreadLocalMap定义所在类），这也意味着只要线程不退出，对象的引用将一直存在。

当线程退出时，Thread类会进行一些清理工作，其中就包括清理ThreadLocalMap，注意下述代码的加粗部分：

```
/**
 * 在线程退出前，由系统回调，进行资源清理
 */
private void exit() {
```

```
    if (group != null) {
        group.threadTerminated(this);
        group = null;
    }
    target = null;
    /* 加速资源清理 */
    threadLocals = null;
    inheritableThreadLocals = null;
    inheritedAccessControlContext = null;
    blocker = null;
    uncaughtExceptionHandler = null;
}
```

因此，如果我们使用线程池，那就意味着当前线程未必会退出（比如固定大小的线程池，线程总是存在）。如果这样，将一些大大的对象设置到ThreadLocal中（它实际保存在线程持有的threadLocals Map内），可能会使系统出现内存泄露的可能（这里我的意思是：你设置了对象到ThreadLocal中，但是不清理它，在你使用几次后，这个对象也不再有用了，但是它却无法被回收）。

此时，如果你希望及时回收对象，最好使用ThreadLocal.remove()方法将这个变量移除。就像我们习惯性地关闭数据库连接一样。如果你确实不需要这个对象了，那么就应该告诉虚拟机，请把它回收掉，防止内存泄露。

另外一种有趣的情况是JDK也可能允许你像释放普通变量一样释放ThreadLocal。比如，我们有时候为了加速垃圾回收，会特意写出类似obj=null之类的代码。如果这么做，obj所指向的对象就会更容易地被垃

圾回收器发现，从而加速回收。

同理，如果对于ThreadLocal的变量，我们也手动将其设置为null，比如tl=null。那么这个ThreadLocal对应的所有线程的局部变量都有可能被回收。这里面的奥秘是什么呢？先来看一个简单的例子：

```
01 public class ThreadLocalDemo_Gc {
02     static volatile ThreadLocal<SimpleDateFormat> tl = new Threa
03         protected void finalize() throws Throwable {
04             System.out.println(this.toString() + " is gc");
05         }
06     };
07     static volatile CountDownLatch cd = new CountDownLatch(100
08     public static class ParseDate implements Runnable {
09         int i = 0;
10         public ParseDate(int i) {
11             this.i = i;
12         }
13         public void run() {
14             try {
15                 if (tl.get() == null) {
16                     tl.set(new SimpleDateFormat("yyyy-MM-dd HH
17                         protected void finalize() throws Throw
18                             System.out.println(this.toString()
19                         }
20                     });
21                 System.out.println(Thread.currentThread().getId())
```

```
22         }
23         Date t = tl.get().parse("2015-03-29 19:29:" +
24     } catch (ParseException e) {
25         e.printStackTrace();
26     } finally {
27         cd.countDown();
28     }
29 }
30 }
31
32 public static void main(String[] args) throws InterruptedException
33     ExecutorService es = Executors.newFixedThreadPool(10);
34     for (int i = 0; i < 10000; i++) {
35         es.execute(new ParseDate(i));
36     }
37     cd.await();
38     System.out.println("mission complete!!");
39     tl = null;
40     System.gc();
41     System.out.println("first GC complete!!");
42     //在设置ThreadLocal的时候，会清除ThreadLocalMap中的无效对象
43     tl = new ThreadLocal<SimpleDateFormat>();
44     cd = new CountDownLatch(10000);
45     for (int i = 0; i < 10000; i++) {
46         es.execute(new ParseDate(i));
47     }
48     cd.await();
```

```
49      Thread.sleep(1000);
50      System.gc();
51      System.out.println("second GC complete!!");
52  }
53 }
```

上述案例是为了跟踪ThreadLocal对象以及内部SimpleDateFormat对象的垃圾回收。为此，我们在第3行和第17行，重载了finalize()方法。这样，我们在对象被回收时，就可以看到它们的踪迹。

在主函数main中，先后进行了两次任务提交，每次10000个任务。在第一次任务提交后，代码第39行，我们将tl设置为null，接着进行一次GC。接着，我们进行第2次任务提交，完成后，在第50行再进行一次GC。

如果你执行上述代码，则最有可能的一种输出如下：

```
10:create SimpleDateFormat
11:create SimpleDateFormat
13:create SimpleDateFormat
17:create SimpleDateFormat
14:create SimpleDateFormat
8:create SimpleDateFormat
16:create SimpleDateFormat
15:create SimpleDateFormat
12:create SimpleDateFormat
9:create SimpleDateFormat
mission complete!!
```

```
first GC complete!!
geym.conc.ch4.tl.ThreadLocalDemo_Gc$1@15f157b is gc
9:create SimpleDateFormat
8:create SimpleDateFormat
16:create SimpleDateFormat
13:create SimpleDateFormat
15:create SimpleDateFormat
10:create SimpleDateFormat
11:create SimpleDateFormat
14:create SimpleDateFormat
17:create SimpleDateFormat
12:create SimpleDateFormat
second GC complete!!
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
geym.conc.ch4.tl.ThreadLocalDemo_Gc$ParseDate$1@4f76f1a0 is gc
```

注意这些输出所代表的含义。首先，线程池中10个线程都各自创建了一个SimpleDateFormat对象实例。接着进行第一次GC，可以看到ThreadLocal对象被回收了（这里使用了匿名类，所以类名看起来有点

怪，这个类就是第2行创建的tl对象）。接着提交了第2次任务，这次一样也创建了10个SimpleDateFormat对象。然后，进行第2次GC。可以看到，在第2次GC后，第一次创建的10个SimpleDateFormat子类实例全部被回收。可以看到，虽然我们没有手工remove()这些对象，但是系统依然有可能回收它们（注意，这段代码是在JDK 7中输出的，在JDK 8中，你也许得不到类似的输出，大家可以比较两个JDK版本之间线程持有ThreadLocal变量的不同）。

要了解这里的回收机制，我们需要更进一步了解Thread.ThreadLocalMap的实现。之前我们说过，ThreadLocalMap是一个类似HashMap的东西。更精确地说，它更加类似WeakHashMap。

ThreadLocalMap的实现使用了弱引用。弱引用是比强引用弱得多的引用。Java虚拟机在垃圾回收时，如果发现弱引用，就会立即回收。ThreadLocalMap内部由一系列Entry构成，每一个Entry都是WeakReference<ThreadLocal>：

```
static class Entry extends WeakReference<ThreadLocal> {
    /** The value associated with this ThreadLocal. */
    Object value;
    Entry(ThreadLocal k, Object v) {
        super(k);
        value = v;
    }
}
```

这里的参数k就是Map的key，v就是Map的value。其中k也就是ThreadLocal实例，作为弱引用使用（super(k)就是调用了WeakReference

的构造函数)。因此，虽然这里使用ThreadLocal作为Map的key，但是实际上，它并不真的持有ThreadLocal的引用。而当ThreadLocal的外部强引用被回收时，ThreadLocalMap中的key就会变成null。当系统进行ThreadLocalMap清理时（比如将新的变量加入表中，就会自动进行一次清理，虽然JDK不一定会进行一次彻底的扫描，但显然在我们这个案例中，它奏效了），就会自然将这些垃圾数据回收。整个结构如图4.1所示。

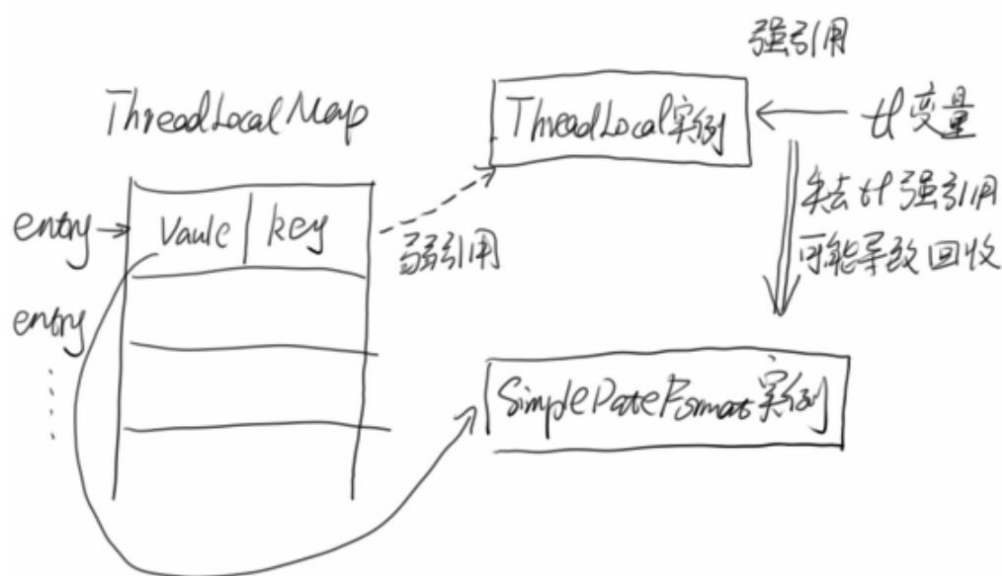


图4.1 ThreadLocal的回收机制

4.3.3 对性能有何帮助

为每一个线程分配一个独立的对象对系统性能也许是有帮助的。当然了，这也不一定，这完全取决于共享对象的内部逻辑。如果共享对象对于竞争的处理容易引起性能损失，我们还是应该考虑使用ThreadLocal为每个线程分配单独的对象。一个典型的案例就是在多线程下产生随机数。

这里，让我们简单测试一下在多线程下产生随机数的性能问题。首先，我们定义一些全局变量：

```
01 public static final int GEN_COUNT = 10000000;  
02 public static final int THREAD_COUNT = 4;  
03 static ExecutorService exe = Executors.newFixedThreadPool(THREAD_COUNT);  
04 public static Random rnd = new Random(123);  
05  
06 public static ThreadLocal<Random> tRnd = new ThreadLocal<Random>() {  
07     @Override  
08     protected Random initialValue() {  
09         return new Random(123);  
10     }  
11 };
```

代码第1行定义了每个线程要产生的随机数数量，第2行定义了参与工作的线程数量，第3行定义了线程池，第4行定义了被多线程共享的Random实例用于产生随机数，第6~11行定义了由ThreadLocal封装的Random。

接着，定义一个工作线程的内部逻辑。它可以工作在两种模式下：

第一是多线程共享一个Random（mode=0），

第二是多个线程各分配一个Random（mode=1）。

```
01 public static class RndTask implements Callable<Long> {  
02     private int mode = 0;  
03 }
```

```
04     public RndTask(int mode) {
05         this.mode = mode;
06     }
07
08     public Random getRandom() {
09         if (mode == 0) {
10             return rnd;
11         } else if (mode == 1) {
12             return tRnd.get();
13         } else {
14             return null;
15         }
16     }
17
18     @Override
19     public Long call() {
20         long b = System.currentTimeMillis();
21         for (long i = 0; i < GEN_COUNT; i++) {
22             getRandom().nextInt();
23         }
24         long e = System.currentTimeMillis();
25         System.out.println(Thread.currentThread().getName() +
26             return e - b;
27     }
28 }
```

上述代码第19~27行定义了线程的工作内容。每个线程会产生若干

个随机数，完成工作后，记录并返回所消耗的时间。

最后是我们的main()函数，它分别对上述两种情况进行测试，并打印了测试的耗时：

```
01 public static void main(String[] args) throws InterruptedException
02     Future<Long>[] futs = new Future[THREAD_COUNT];
03     for (int i = 0; i < THREAD_COUNT; i++) {
04         futs[i] = exe.submit(new RndTask(0));
05     }
06     long totaltime = 0;
07     for (int i = 0; i < THREAD_COUNT; i++) {
08         totaltime += futs[i].get();
09     }
10     System.out.println("多线程访问同一个Random实例:" + totaltime);
11
12     //ThreadLocal的情况
13     for (int i = 0; i < THREAD_COUNT; i++) {
14         futs[i] = exe.submit(new RndTask(1));
15     }
16     totaltime = 0;
17     for (int i = 0; i < THREAD_COUNT; i++) {
18         totaltime += futs[i].get();
19     }
20     System.out.println("使用ThreadLocal包装Random实例:" + totaltime);
21     exe.shutdown();
22 }
```

上述代码的运行结果，可能如下：

```
pool-1-thread-3 spend 3398ms
pool-1-thread-1 spend 3436ms
pool-1-thread-2 spend 3495ms
pool-1-thread-4 spend 3513ms
多线程访问同一个Random实例:13842ms
pool-1-thread-4 spend 375ms
pool-1-thread-1 spend 429ms
pool-1-thread-2 spend 453ms
pool-1-thread-3 spend 499ms
使用ThreadLocal包装Random实例:1756ms
```

很明显，在多线程共享一个Random实例的情况下，总耗时达13秒之多（这里是指4个线程的耗时总和，不是程序执行的经历时间）。而在ThreadLocal模式下，仅耗时1.7秒左右。

4.4 无锁

就人的性格而言，我们可以分为乐天派和悲观派。对于乐天派来说，总是会把事情往好的方面想。他们认为所有事情总是不太容易发现问题，出错是小概率的，所以我们可以肆无忌惮地做事。如果真的不幸遇到了问题，则有则改之无则加勉。而对于悲观的人群来说，他们总是担惊受怕，认为出错是一种常态，所以无论巨细，都考虑得面面俱到，滴水不漏，确保为人处世，万无一失。

对于并发控制而言，锁是一种悲观的策略。它总是假设每一次的临界区操作会产生冲突，因此，必须对每次操作都小心翼翼。如果有多个线程同时需要访问临界区资源，就宁可牺牲性能让线程进行等待，所以说锁会阻塞线程执行。而无锁是一种乐观的策略，它会假设对资源的访问是没有冲突的。既然没有冲突，自然不需要等待，所以所有的线程都可以在不停顿的状态下持续执行。那遇到冲突怎么办呢？无锁的策略使用一种叫做比较交换的技术（CAS Compare And Swap）来鉴别线程冲突，一旦检测到冲突产生，就重试当前操作直到没有冲突为止。

4.4.1 与众不同的并发策略：比较交换（CAS）

与锁相比，使用比较交换（下文简称CAS）会使程序看起来更加复杂一些。但由于其非阻塞性，它对死锁问题天生免疫，并且，线程间的相互影响也远远比基于锁的方式要小。更为重要的是，使用无锁的方式

完全没有锁竞争带来的系统开销，也没有线程间频繁调度带来的开销，因此，它要比基于锁的方式拥有更优越的性能。

CAS算法的过程是这样：它包含三个参数CAS(V,E,N)。V表示要更新的变量，E表示预期值，N表示新值。仅当V值等于E值时，才会将V的值设为N，如果V值和E值不同，则说明已经有其他线程做了更新，则当前线程什么都不做。最后，CAS返回当前V的真实值。CAS操作是抱着乐观的态度进行的，它总是认为自己可以成功完成操作。当多个线程同时使用CAS操作一个变量时，只有一个会胜出，并成功更新，其余均会失败。失败的线程不会被挂起，仅是被告知失败，并且允许再次尝试，当然也允许失败的线程放弃操作。基于这样的原理，CAS操作即使没有锁，也可以发现其他线程对当前线程的干扰，并进行恰当的处理。

简单地说，CAS需要你额外给出一个期望值，也就是你认为这个变量现在应该是什么样子的。如果变量不是你想象的那样，那说明它已经被别人修改过了。你就重新读取，再次尝试修改就好了。

在硬件层面，大部分的现代处理器都已经支持原子化的CAS指令。在JDK 5.0以后，虚拟机便可以使用这个指令来实现并发操作和并发数据结构，并且，这种操作在虚拟机中可以说是无处不在。

4.4.2 无锁的线程安全整数： **AtomicInteger**

为了让Java程序员能够受益于CAS等CPU指令，JDK并发包中有一个atomic包，里面实现了一些直接使用CAS操作的线程安全的类型。

其中，最常用的一个类，应该就是AtomicInteger。你可以把它看做是一个整数。但是与Integer不同，它是可变的，并且是线程安全的。对其进行修改等任何操作，都是用CAS指令进行的。这里简单列举一下AtomicInteger的一些主要方法，对于其他原子类，操作也是非常类似的：

```
public final int get() //取得当前值
public final void set(int newValue) //设置当前值
public final int getAndSet(int newValue) //设置新值，
public final boolean compareAndSet(int expect, int u) //如果当前值
public final int getAndIncrement() //当前值加1,
public final int getAndDecrement() //当前值减1,
public final int getAndAdd(int delta) //当前值增加
public final int incrementAndGet() //当前值加1,
public final int decrementAndGet() //当前值减1,
public final int addAndGet(int delta) //当前值增加
```

就内部实现上来说，AtomicInteger中保存一个核心字段：

```
private volatile int value;
```

它就代表了AtomicInteger的当前实际取值。此外还有一个：

```
private static final long valueOffset;
```

它保存着value字段在AtomicInteger对象中的偏移量。后面你会看到，这个偏移量是实现AtomicInteger的关键。

AtomicInteger的使用非常简单，这里给出一个示例：

```
01 public class AtomicIntegerDemo {
02     static AtomicInteger i=new AtomicInteger();
03     public static class AddThread implements Runnable{
04         public void run(){
05             for(int k=0;k<10000;k++)
06                 i.incrementAndGet();
07         }
08     }
09     public static void main(String[] args) throws InterruptedException
10         Thread[] ts=new Thread[10];
11         for(int k=0;k<10;k++){
12             ts[k]=new Thread(new AddThread());
13         }
14         for(int k=0;k<10;k++){ts[k].start();}
15         for(int k=0;k<10;k++){ts[k].join();}
16         System.out.println(i);
17     }
18 }
```

第6行的`AtomicInteger.incrementAndGet()`方法会使用CAS操作将自己加1，同时也会返回当前值（这里忽略了当前值）。如果你执行这段代码，你会看到程序输出了100000。这说明程序正常执行，没有错误。如果不是线程安全，i的值应该会小于100000才对。

使用`AtomicInteger`会比使用锁具有更好的性能。出于篇幅限制，这里不再给出`AtomicInteger`和锁的性能对比的测试代码，相信写一段简单的小代码测试两者的性能应该不是难事。这里让我们关注一下

incrementAndGet()的内部实现（我们基于JDK 1.7分析，JDK 1.8与1.7的实现有所不同）。

```
1 public final int incrementAndGet() {  
2     for (;;) {  
3         int current = get();  
4         int next = current + 1;  
5         if (compareAndSet(current, next))  
6             return next;  
7     }  
8 }
```

其中get()方法非常简单，就是返回内部数据value。

```
public final int get() {  
    return value;  
}
```

这里让人映像深刻的，应该是incrementAndGet()方法的第2行for循环吧！如果你是初次看到这样的代码，可能会觉得很奇怪，为什么连设置一个值那么简单的操作都需要一个死循环呢？原因就是：CAS操作未必是成功的，因此对于不成功的情况，我们就需要进行不断的尝试。第3行的get()取得当前值，接着加1后得到新值next。这里，我们就得到了CAS必需的两个参数：期望值以及新值。使用compareAndSet()方法将新值next写入，成功的条件是在写入的时刻，当前的值应该要等于刚刚取得的current。如果不是这样，就说明AtomicInteger的值在第3行到第5行代码之间，又被其他线程修改过了。当前线程看到的状态就是一个过期状态。因此，compareAndSet返回失败，需要进行下一次重试，直到成

功。

以上就是CAS操作的基本思想。在后面我们会看到，无论程序多么复杂，其基本原理总是不变的。

和AtomicInteger类似的类还有AtomicLong用来代表long型，AtomicBoolean表示boolean型，AtomicReference表示对象引用。

4.4.3 Java中的指针：Unsafe类

如果你对技术有着不屈不挠的追求，应该还会特别在意incrementAndGet()方法中compareAndSet()的实现。现在，就让我们更进一步看一下它吧！

```
public final boolean compareAndSet(int expect, int update) {  
    return unsafe.compareAndSwapInt(this, valueOffset, expect, up  
}
```

在这里，我们看到一个特殊的变量unsafe，它是sun.misc.Unsafe类型。从名字看，这个类应该是封装了一些不安全的操作。那什么操作是不安全的呢？学习过C或者C++的话，大家应该知道，指针是不安全的，这也是在Java中把指针去除的重要原因。如果指针指错了位置，或者计算指针偏移量时出错，结果可能是灾难性的，你很有可能会覆盖别人的内存，导致系统崩溃。

而这里的Unsafe就是封装了一些类似指针的操作。compareAndSwapInt()方法是一个native方法，它的几个参数含义如下：

```
public final native boolean compareAndSwapInt(Object o, long offs
```

第一个参数`o`为给定的对象，`offset`为对象内的偏移量（其实就是一个字段到对象头部的偏移量，通过这个偏移量可以快速定位字段），`expected`表示期望值，`x`表示要设置的值。如果指定的字段的值等于`expected`，那么就会把它设置为`x`。

不难看出，`compareAndSwapInt()`方法的内部，必然是使用CAS原子指令来完成的。此外，`Unsafe`类还提供了一些方法，主要有以下几个（以`Int`操作为例，其他数据类型是类似的）：

```
//获得给定对象偏移量上的int值
public native int getInt(Object o, long offset);
//设置给定对象偏移量上的int值
public native void putInt(Object o, long offset, int x);
//获得字段在对象中的偏移量
public native long objectFieldOffset(Field f);
//设置给定对象的int值，使用volatile语义
public native void putIntVolatile(Object o, long offset, int x);
//获得给定对象对象的int值，使用volatile语义
public native int      getIntVolatile(Object o, long offset);
//和putIntVolatile()一样，但是它要求被操作字段就是volatile类型的
public native void putOrderedInt(Object o, long offset, int x);
```

如果大家还记得“3.3.4深度剖析`ConcurrentLinkedQueue`”一节中描述的`ConcurrentLinkedQueue`实现，应该对`ConcurrentLinkedQueue`中的`Node`还有些印象。`Node`的一些CAS操作也都是使用`Unsafe`类来实现的。大家可以回顾一下，以加深对`Unsafe`类的印象。

这里就可以看到，虽然Java抛弃了指针。但是在关键时刻，类似指针的技术还是必不可少的。这里底层的Unsafe实现就是最好的例子。但是很不幸，JDK的开发人员并不希望大家使用这个类。获得Unsafe实例的方法是调动其工厂方法getUnsafe()。但是，它的实现却是这样：

```
public static Unsafe getUnsafe() {  
    Class cc = Reflection.getCallerClass();  
    if (cc.getClassLoader() != null)  
        throw new SecurityException("Unsafe");  
    return theUnsafe;  
}
```

注意加粗部分的代码，它会检查调用getUnsafe()函数的类，如果这个类的ClassLoader不为null，就直接抛出异常，拒绝工作。因此，这也使得我们自己的应用程序无法直接使用Unsafe类。它是一个JDK内部使用的专属类。

注意：根据Java类加载器的工作原理，应用程序的类由App Loader加载。而系统核心类，如rt.jar中的类由Bootstrap类加载器加载。Bootstrap加载器没有Java对象的对象，因此试图获得这个类加载器会返回null。所以，当一个类的类加载器为null时，说明它是由Bootstrap加载的，而这个类也极有可能是rt.jar中的类。

4.4.4 无锁的对象引用： AtomicReference

AtomicReference和AtomicInteger非常类似，不同之处就在于

`AtomicInteger`是对整数的封装，而`AtomicReference`则对应普通的对象引用。也就是它可以保证你在修改对象引用时的线程安全性。在介绍`AtomicReference`的同时，我希望同时提出一个有关原子操作的逻辑上的不足。

之前我们说过，线程判断被修改对象是否可以正确写入的条件是对象的当前值和期望值是否一致。这个逻辑从一般意义上来说是正确的。但有可能出现一个小小的例外，就是当你获得对象当前数据后，在准备修改为新值前，对象的值被其他线程连续修改了两次，而经过这两次修改后，对象的值又恢复为旧值。这样，当前线程就无法正确判断这个对象究竟是否被修改过。如图4.2所示，显示了这种情况。

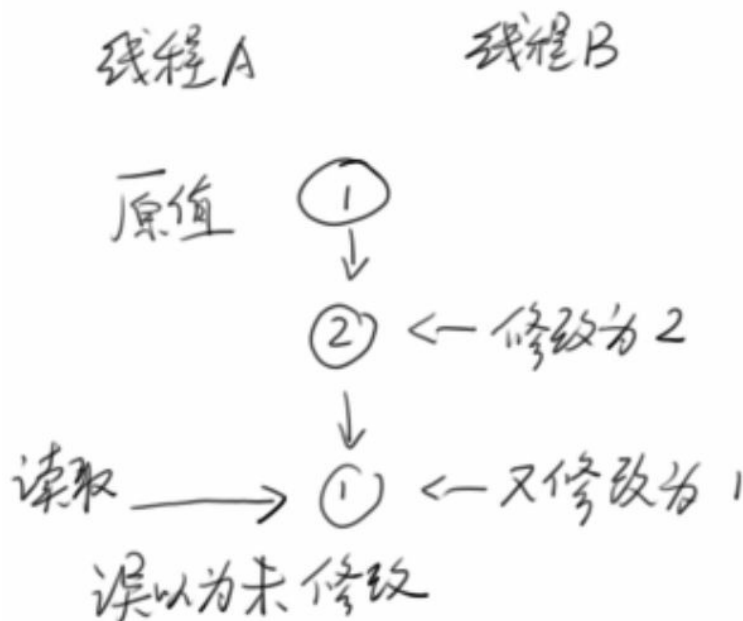


图4.2 对象值被反复修改回原数据

一般来说，发生这种情况的概率很小。而且即使发生了，可能也不是什么大问题。比如，我们只是简单地要做一个数值加法，即使在我取得期望值后，这个数字被不断的修改，只要它最终改回了我的期望值，我的加法计算就不会出错。也就是说，当你修改的对象没有过程的状态

信息，所有的信息都只保存于对象的数值本身。

但是，在现实中，还可能存在另外一种场景，就是我们是否能修改对象的值，不仅取决于当前值，还和对象的过程变化有关，这时，AtomicReference就无能为力了。

打一个比方，如果有一家蛋糕店，为了挽留客户，决定为贵宾卡里余额小于20元的客户一次性赠送20元，刺激消费者充值和消费。但条件是，每一位客户只能被赠送一次。

现在，我们就来模拟这个场景，为了演示AtomicReference，我在这里使用AtomicReference实现这个功能。首先，我们模拟用户账户余额。

定义用户账户余额：

```
static AtomicReference<Integer> money=new AtomicReference<Integer>()
// 设置账户初始值小于20，显然这是一个需要被充值的账户
money.set(19);
```

接着，我们需要若干个后台线程，它们不断扫描数据，并为满足条件的客户充值。

```
01 //模拟多个线程同时更新后台数据库，为用户充值
02 for(int i = 0 ; i < 3 ; i++) {
03     new Thread() {
04         public void run() {
05             while(true){
06                 while(true){
07                     Integer m=money.get();
```

```

08             if(m<20){
09                 if(money.compareAndSet(m, m+20)){
10                     System.out.println("余额小于20元，充值成功，余额
11                         break;
12                 }
13             }else{
14                 //System.out.println("余额大于20元，无须充
15                 break ;
16             }
17         }
18     }
19 }
20     }.start();
21 }

```

上述代码第8行，判断用户余额并给予赠送金额。如果已经被其他用户处理，那么当前线程就会失败。因此，可以确保用户只会被充值一次。

此时，如果很不幸，用户正好正在进行消费，就在赠予金额到账的同时，他进行了一次消费，使得总金额又小于20元，并且正好累计消费了20元。使得消费、赠予后的金额等于消费前、赠予前的金额。这时，后台的赠予进程就会误以为这个账户还没有赠予，所以，存在被多次赠予的可能。下面模拟了这个消费线程：

```

01 //用户消费线程，模拟消费行为
02 new Thread() {
03     public void run() {

```

```
04         for(int i=0;i<100;i++){
05             while(true){
06                 Integer m=money.get();
07                 if(m>10){
08                     System.out.println("大于10元");
09                     if(money.compareAndSet(m, m-10)){
10                         System.out.println("成功消费10元，余额："+
11                             break;
12                     }
13                 }else{
14                     System.out.println("没有足够的金额");
15                     break;
16                 }
17             }
18             try {Thread.sleep(100);} catch (InterruptedException
19         }
20     }
21 }.start();
```

上述代码中，消费者只要贵宾卡里的钱大于10元，就会立即进行一次10元的消费。执行上述程序，得到的输出如下：

余额小于20元，充值成功，余额:39元

大于10元

成功消费10元，余额:29

大于10元

成功消费10元，余额:19

余额小于20元，充值成功，余额:39元

大于10元

成功消费10元，余额:29

大于10元

成功消费10元，余额:39

余额小于20元，充值成功，余额:39元

从这一段输出中，可以看到，这个账户被先后反复多次充值。其原因正是因为账户余额被反复修改，修改后的值等于原有的数值，使得CAS操作无法正确判断当前数据状态。

虽然说这种情况出现的概率不大，但是依然是有可能出现的。因此，当业务上确实可能出现这种情况时，我们也必须多加防范。体贴的JDK也已经为我们考虑到了这种情况，使用AtomicStampedReference就可以很好地解决这个问题。

4.4.5 带有时间戳的对象引用： AtomicStampedReference

AtomicReference无法解决上述问题的根本因为是对象在修改过程中，丢失了状态信息。对象值本身与状态被画上了等号。因此，我们只要能够记录对象在修改过程中的状态值，就可以很好地解决对象被反复修改导致线程无法正确判断对象状态的问题。

AtomicStampedReference正是这么做的。它内部不仅维护了对象值，还维护了一个时间戳（我这里把它称为时间戳，实际上它可以使任何一个整数来表示状态值）。当AtomicStampedReference对应的数值被

修改时，除了更新数据本身外，还必须要更新时间戳。当AtomicStampedReference设置对象值时，对象值以及时间戳都必须满足期望值，写入才会成功。因此，即使对象值被反复读写，写回原值，只要时间戳发生变化，就能防止不恰当的写入。

AtomicStampedReference的几个API在AtomicReference的基础上新增了有关时间戳的信息：

```
//比较设置 参数依次为：期望值 写入新值 期望时间戳 新时间戳
public boolean compareAndSet(V expectedReference,V
newReference,int expectedStamp,int newStamp)
//获得当前对象引用
public V getReference()
//获得当前时间戳
public int getStamp()
//设置当前对象引用和时间戳
public void set(V newReference, int newStamp)
```

有了AtomicStampedReference这个法宝，我们就再也不用担心对象被写坏啦！现在，就让我们使用AtomicStampedReference来修正那个贵宾卡充值的问题：

```
01 public class AtomicStampedReferenceDemo {
02     static AtomicStampedReference<Integer> money=new AtomicStampe
03     public static void main(String[] args) {
04         //模拟多个线程同时更新后台数据库，为用户充值
05         for(int i = 0 ; i < 3 ; i++) {
06             final int timestamp=money.getStamp();
```

```
07         new Thread() {
08             public void run() {
09                 while(true){
10                     while(true){
11                         Integer m=money.getReference();
12                         if(m<20){
13                             if(money.compareAndSet(m, m+20,timest
14 System.out.println("余额小于20元, 充值成功, 余额:"+money
15                             break;
16                         }
17                     }else{
18                         //System.out.println("余额大于20
19                         break ;
20                     }
21                 }
22             }
23         }
24     }.start();
25 }
26
27 //用户消费线程, 模拟消费行为
28 new Thread() {
29     public void run() {
30         for(int i=0;i<100;i++){
31             while(true){
32                 int timestamp=money.getStamp();
33                 Integer m=money.getReference();
```

```

34             if(m>10){
35                 System.out.println("大于10元");
36                 if(money.compareAndSet(m, m-10,times
37             System.out.println("成功消费10元，余额："+m
38                 break;
39             }
40         }else{
41             System.out.println("没有足够的金额");
42             break;
43         }
44     }
45     try {Thread.sleep(100);} catch (Interrupte
46     }
47 }
48 }.start();
49 }
50 }

```

第2行，我们使用AtomicStampedReference代替原来的AtomicReference。第6行获得账户的时间戳，后续的赠予操作以这个时间戳为依据。如果赠予成功（第13行），则修改时间戳，使得系统不可能发生二次赠予的情况。消费线程也是类似，每次操作，都使得时间戳加1（第36行），使之不可能重复。

执行上述代码，可以得到以下输出：

```

余额小于20元，充值成功，余额:39元
大于10元

```

成功消费10元，余额:29

大于10元

成功消费10元，余额:19

大于10元

成功消费10元，余额:9

没有足够的金额

可以看到，账户只被赠予了一次。

4.4.6 数组也能无锁： **AtomicIntegerArray**

除了提供基本数据类型外，JDK还为我们准备了数组等复合结构。当前可用的原子数组有：`AtomicIntegerArray`、`AtomicLongArray`和`AtomicReferenceArray`，分别表示整数数组、`long`型数组和普通的对象数组。

这里以`AtomicIntegerArray`为例，展示原子数组的使用方式。

`AtomicIntegerArray`本质上是对`int[]`类型的封装，使用`Unsafe`类通过CAS的方式控制`int[]`在多线程下的安全性。它提供了以下几个核心API：

```
//获得数组第i个下标的元素
public final int get(int i)
//获得数组的长度
public final int length()
```

```

//将数组第i个下标设置为newValue，并返回旧的值
public final int getAndSet(int i, int newValue)
//进行CAS操作，如果第i个下标的元素等于expect，则设置为update，设置成功返回1
public final boolean compareAndSet(int i, int expect, int update)
//将第i个下标的元素加1
public final int getAndIncrement(int i)
//将第i个下标的元素减1
public final int getAndDecrement(int i)
//将第i个下标的元素增加delta（delta可以是负数）
public final int getAndAdd(int i, int delta)

```

下面给出一个简单的示例，展示AtomicIntegerArray的使用：

```

01 public class AtomicIntegerArrayDemo {
02     static AtomicIntegerArray arr = new AtomicIntegerArray(10)
03     public static class AddThread implements Runnable{
04         public void run(){
05             for(int k=0;k<10000;k++){
06                 arr.getAndIncrement(k%arr.length());
07             }
08         }
09     public static void main(String[] args) throws InterruptedException{
10         Thread[] ts=new Thread[10];
11         for(int k=0;k<10;k++){
12             ts[k]=new Thread(new AddThread());
13         }
14         for(int k=0;k<10;k++){ts[k].start();}

```

```
15         for(int k=0;k<10;k++){ts[k].join();}  
16         System.out.println(arr);  
17     }  
18 }
```

上述代码第2行，申明了一个内含10个元素的数组。第3行定义的线程对数组内10个元素进行累加操作，每个元素各加1000次。第11行，开启10个这样的线程。因此，可以预测，如果线程安全，数组内10个元素的值必然都是10000。反之，如果线程不安全，则部分或者全部数值会小于10000。

程序的输出结果如下：

```
[10000, 10000, 10000, 10000, 10000, 10000, 10000, 10000, 10000, 1
```

这说明AtomicIntegerArray确实合理地保证了数组的线程安全性。

4.4.7 让普通变量也享受原子操作： AtomicIntegerFieldUpdater

有时候，由于初期考虑不周，或者后期的需求变化，一些普通变量可能也会有线程安全的需求。如果改动不大，我们可以简单地修改程序中每一个使用或者读取这个变量的地方。但显然，这样并不符合软件设计中的一条重要原则——开闭原则。也就是系统对功能的增加应该是开放的，而对修改应该是相对保守的。而且，如果系统里使用到这个变量的地方特别多，一个一个修改也是一件令人厌烦的事情（况且很多使用场景下可能只是只读的，并无线程安全的强烈要求，完全可以保持原

样)。

如果你有这种困扰，在这里根本不需要担心，因为在原子包里还有一个实用的工具类AtomicIntegerFieldUpdater。它可以让你在不改动（或者极少改动）原有代码的基础上，让普通的变量也享受CAS操作带来的线程安全性，这样你可以修改极少的代码，来获得线程安全的保证。这听起来是不是让人很激动呢？

根据数据类型不同，这个Updater有三种，分别是AtomicIntegerFieldUpdater、AtomicLongFieldUpdater和AtomicReferenceFieldUpdater。顾名思义，它们分别可以对int、long和普通对象进行CAS修改。

现在来思考这么一个场景。假设某地要进行一次选举。现在模拟这个投票场景，如果选民投了候选人一票，就记为1，否则记为0。最终的选票显然就是所有数据的简单求和。

```
01 public class AtomicIntegerFieldUpdaterDemo {
02     public static class Candidate{
03         int id;
04         volatile int score;
05     }
06     public final static AtomicIntegerFieldUpdater<Candidate>
07         = AtomicIntegerFieldUpdater.newUpdater(Candidate.class
08         //检查Updater是否工作正确
09     public static AtomicInteger allScore=new AtomicInteger(0);
10     public static void main(String[] args) throws InterruptedException
11         final Candidate stu=new Candidate();
```



```

12      Thread[] t=new Thread[10000];
13      for(int i = 0 ; i < 10000 ; i++) {
14          t[i]=new Thread() {
15              public void run() {
16                  if(Math.random()>0.4){
17                      scoreUpdater.incrementAndGet(stu);
18                      allScore.incrementAndGet();
19                  }
20              }
21          };
22          t[i].start();
23      }
24      for(int i = 0 ; i < 10000 ; i++) { t[i].join();}
25      System.out.println("score="+stu.score);
26      System.out.println("allScore="+allScore);
27  }
28 }

```

上述代码模拟了这个计票场景，候选人的得票数量记录在Candidate.score中。注意，它是一个普通的volatile变量。而volatile变量并不是线程安全的。第6~7行定义了AtomicIntegerFieldUpdater实例，用来对Candidate.score进行写入。而后续的allScore我们用来检查AtomicIntegerFieldUpdater的正确性。如果AtomicIntegerFieldUpdater真的保证了线程安全，那么最终Candidate.score和allScore的值必然是相等的。否则，就说明AtomicIntegerFieldUpdater根本没有确保线程安全的写入。第12~21行模拟了计票过程，这里假设有大约60%的人投赞成票，并且投票是随机进行的。第17行使用Updater修改Candidate.score（这里

应该是线程安全的），第18行使用AtomicInteger计数，作为参考基准。

大家如果运行这段程序，不难发现，最终的Candidate.score总是和allScore绝对相等。这说明AtomicIntegerFieldUpdater很好地保证了Candidate.score的线程安全。

虽然AtomicIntegerFieldUpdater很好用，但是还是有几个注意事项：

第一，Updater只能修改它可见范围内的变量。因为Updater使用反射得到这个变量。如果变量不可见，就会出错。比如如果score申明为private，就是不可行的。

第二，为了确保变量被正确的读取，它必须是volatile类型的。如果我们原有代码中未申明这个类型，那么简单地申明一下就行，这不会引起什么问题。

第三，由于CAS操作会通过对象实例中的偏移量直接进行赋值，因此，它不支持static字段（Unsafe.objectFieldOffset()不支持静态变量）。

好了，通过AtomicIntegerFieldUpdater，是不是让我们可以更加随心所欲地对系统关键数据进行线程安全的保护呢？

4.4.8 挑战无锁算法：无锁的Vector实现

我们已经比较完整地介绍了有关无锁的概念和使用方法。相对于有锁的方法，使用无锁的方式编程更加考验一个程序员的耐心和智力。但是，无锁带来的好处也是显而易见的，第一，在高并发的情况下，它比

有锁的程序拥有更好的性能；第二，它天生就是死锁免疫的。就凭借这两个优势，就值得我们冒险尝试使用无锁的并发。

这里，我想向大家介绍一种使用无锁方式实现的Vector。通过这个案例，我们可以更加深刻地认识无锁的算法，同时也可以学习一下有关Vector实现的细节和算法技巧（在本例中，讲述的无锁Vector来自于amino并发包）。

我们将这个无锁的Vector称为LockFreeVector。它的特点是可以根据需求动态扩展其内部空间。在这里，我们使用二维数组来表示LockFreeVector的内部存储，如下：

```
private final AtomicReferenceArray<AtomicReferenceArray<E>> buckets;
```

变量buckets存放所有的内部元素。从定义上看，它是一个保存着数组的数组，也就是通常的二维数组。特别之处在于这些数组都是使用CAS的原子数组。为什么使用二维数组去实现一个一维的Vector呢？这是为了将来Vector进行动态扩展时可以更加方便。我们知道，AtomicReferenceArray内部使用Object[]来进行实际数据的存储，这使得动态空间增加特别的麻烦，因此使用二维数组的好处就是为了将来可以方便地增加新的元素。

此外，为了更有序的读写数组，定义一个称为Descriptor的元素。它的作用是使用CAS操作写入新数据。

```
01 static class Descriptor<E> {
02     public int size;
03     volatile WriteDescriptor<E> writeop;
04     public Descriptor(int size, WriteDescriptor<E> writeop) {
```

```
05         this.size = size;
06         this.writeop = writeop;
07     }
08     public void completeWrite() {
09         WriteDescriptor<E> tmpOp = writeop;
10         if (tmpOp != null) {
11             tmpOp.doIt();
12             writeop = null; // this is safe since all write to
13                             // null as r_value.
14         }
15     }
16 }
17
18 static class WriteDescriptor<E> {
19     public E oldV;
20     public E newV;
21     public AtomicReferenceArray<E> addr;
22     public int addr_ind;
23
24     public WriteDescriptor(AtomicReferenceArray<E> addr, int
25                             E oldV, E newV) {
26         this.addr = addr;
27         this.addr_ind = addr_ind;
28         this.oldV = oldV;
29         this.newV = newV;
30     }
31
```

```

32     public void doIt() {
33         addr.compareAndSet(addr_ind, oldV, newV);
34     }
35 }

```

上述代码第4行定义的Descriptor构造函数接收两个参数，第一个为整个Vector的长度，第2个为一个writer。最终，写入数据是通过writer进行的（通过completeWrite()方法）。

第24行，WriteDescriptor的构造函数接收四个参数。第一个参数addr表示要修改的原子数组，第二个参数为要写入的数组索引位置，第三个oldV为期望值，第四个newV为需要写入的值。

在构造LockFreeVector时，显然需要将buckets和descriptor进行初始化。

```

public LockFreeVector() {
    buckets = new AtomicReferenceArray<AtomicReferenceArray<E>>
    buckets.set(0, new AtomicReferenceArray<E>(FIRST_BUCKET_SIZE
    descriptor = new AtomicReference<Descriptor<E>>(new Descriptor
        null));
}

```

在这里N_BUCKET为30,也就是说这个buckets里面可以存放一共30个数组（由于数组无法动态增长，因此数组总数也就不能超过30个）。并且将第一个数组的大小FIRST_BUCKET_SIZE设为8。到这里，大家可能会有一个疑问，如果每个数组8个元素，一共30个数组，那岂不是一共只能存放240个元素吗？

如果大家了解JDK内的Vector实现，应该知道，Vector在进行空间增长时，默认情况下，每次都会将总容量翻倍。因此，这里也借鉴类似的思想，每次空间扩张，新的数组的大小为原来的两倍（即每次空间扩展都启用一个新的数组），因此，第一个数组为8，第二个就是16，第三个就是32。依此类推，因此30个数组可以支持的总元素达到 2^{30} 。

这数值已经超过了 2^{33} ，即在80亿以上。因此，可以满足一般的应用。

当有元素需要加入LockFreeVector时，使用一个名为push_back()的方法，将元素压入Vector最后一个位置。这个操作显然就是LockFreeVector的最为核心的方法，也是最能体现CAS使用特点的方法，它的实现如下：

```
01 public void push_back(E e) {
02     Descriptor<E> desc;
03     Descriptor<E> newd;
04     do {
05         desc = descriptor.get();
06         desc.completeWrite();
07
08         int pos = desc.size + FIRST_BUCKET_SIZE;
09         int zeroNumPos = Integer.numberOfLeadingZeros(pos);
10         int bucketInd = zeroNumFirst - zeroNumPos;
11         if (buckets.get(bucketInd) == null) {
12             int newLen = 2 * buckets.get(bucketInd - 1).length
13             if (debug)
14                 System.out.println("New Length is:" + newLen);
```

```

15         buckets.compareAndSet(bucketInd, null,
16             new AtomicReferenceArray<E>(newLen));
17     }
18
19     int idx = (0x80000000 >>> zeroNumPos) ^ pos;
20     newd = new Descriptor<E>(desc.size + 1, new WriteDesc
21         buckets.get(bucketInd), idx, null, e));
22     } while (!descriptor.compareAndSet(desc, newd));
23     descriptor.get().completeWrite();
24 }

```

可以看到，这个方法主体部分是一个do-while循环，用来不断尝试对descriptor的设置。也就是通过CAS保证了descriptor的一致性和安全性。在第23行，使用descriptor将数据真正地写入数组中。这个descriptor写入的数据由第20～21行构造的WriteDescriptor决定。

在循环最开始（第5行），使用descriptor先将数据写入数组，是为了防止上一个线程设置完descriptor后（第22行），还没来得及执行第23行的写入，因此，做一次预防性的操作。

因为限制要将元素e压入Vector，因此，我们必须首先知道这个e应该放在哪个位置。由于目前使用了二维数组，因此我们自然需要知道e所在哪个数组（buckets中的下标位置）和数组中的下标。

第8～10行通过当前Vector的大小（desc.size），计算新的元素应该落入哪个数组。这里使用了位运算进行计算。

之前说过，LockFreeVector每次都会成倍的扩容。它的第1个数组长

度为8，第2个就是16，第3个就是32，依此类推。它们的二进制表示如下。

- 00000000 00000000 00000000 00001000：第一个数组大小，28个前导零。
- 00000000 00000000 00000000 00010000：第二个数组大小，27个前导零。
- 00000000 00000000 00000000 00100000：第三个数组大小，26个前导零。
- 00000000 00000000 00000000 01000000：第四个数组大小，25个前导零。

它们之和就是整个LockFreeVector的总大小，因此，如果每一个数组都恰好填满，那么总大小应该类似如下的数值（以4个数组填满为例）。

- 00000000 00000000 00000000 01111000：4个数组都恰好填满时的大小。

导致这个数字进位的最小条件，就是加上二进制的1000。而这个数字正好是8（FIRST_BUCKET_SIZE就是8）。这就是第8行代码的意义。它可以使得数组大小发生一次二进制的进位（如果不进位说明还在第一个数组中），进位后前导零的数量就会发生变化。而元素所在的数组，和pos（第8行定义的变量）的前导零直接相关。每进行一次数组扩容，它的前导零就会减1。如果从来没有扩容过，它的前导零就是28个。以后，逐级减1。这就是第9行获得pos前导零的原因。第10行，通过pos的前导零可以立即定位使用哪个数组（也就是得到了bucketInd的值）。

第11行，判断这个数组是否存在。如果不存在，则创建这个数组，大小为前一个数组的两倍，并把它设置到buckets中。

接着再看一下元素没有恰好填满的情况。

- 00000000 00000000 00000000 00001000: 第一个数组大小，28个前导零。
- 00000000 00000000 00000000 00010000: 第二个数组大小，27个前导零。
- 00000000 00000000 00000000 00100000: 第三个数组大小，26个前导零。
- 00000000 00000000 00000000 00000001: 第四个数组大小，只有一个元素。

那么总大小如下。

- 00000000 00000000 00000000 00111001: 元素总个数。

总个数加上二进制1000后，得到：

- 00000000 00000000 00000000 01000001

显然，通过前导零可以定位到第4个数组。而剩余位，显然就表示元素在当前数组内的偏移量（也就是数组下标）。根据这个理论，我们就可以通过pos计算这个元素应该放在给定数组的哪个位置。通过第19行代码，获得pos的除了第一位数字1以外的其他位的数值。因此，pos的前导零可以表示元素所在的数组，而pos的后面几位，则表示元素在这个数组中的位置。由此，第19行代码就取得了元素的所在位置idx。

到此，我们就已经得到新元素位置的全部信息，剩下的就是将这些信息传递给Descriptor让它在给定的位置把元素e安置上去即可。这里，就通过CAS操作，保证写入正确性。

下面来看一下get()操作的实现：

```
1 @Override
2 public E get(int index) {
3     int pos = index + FIRST_BUCKET_SIZE;
4     int zeroNumPos = Integer.numberOfLeadingZeros(pos);
5     int bucketInd = zeroNumFirst - zeroNumPos;
6     int idx = (0x80000000 >>> zeroNumPos) ^ pos;
7     return buckets.get(bucketInd).get(idx);
8 }
```

在get()的实现中，第3~6行使用了相同的算法获得所需元素的数组以及数组中的索引下标。这里简单地通过buckets定位到对应的元素即可。

这样，对于Vector来说两个重要的方法就已经实现了。其他方法也是非常类似的，这里就不再详细讨论了。

4.4.9 让线程之间互相帮助：细看SynchronousQueue的实现

在对线程池的介绍中，提到了一个非常特殊的等待队列SynchronousQueue。SynchronousQueue的容量为0，任何一个对

SynchronousQueue的写需要等待一个对SynchronousQueue的读，反之亦然。因此，SynchronousQueue与其说是一个队列，不如说是一个数据交换通道。那SynchronousQueue的奇妙功能是如何实现的呢？

既然我打算在这一节中介绍它，那么SynchronousQueue就和无锁的操作脱离不了关系。实际上SynchronousQueue内部也正是大量使用了无锁工具。

对SynchronousQueue来说，它将put()和take()两个功能截然不同的操作抽象为一个共通的方法Transferer.transfer()。从字面上看，这就是数据传递的意思。它的完整签名如下：

```
Object transfer(Object e, boolean timed, long nanos)
```

当参数e为非空时，表示当前操作传递给一个消费者，如果为空，则表示当前操作需要请求一个数据。timed参数决定是否存在timeout时间，nanos决定了timeout的时长。如果返回值非空，则表示数据已经接受或者正常提供，如果为空，则表示失败（超时或者中断）。

SynchronousQueue内部会维护一个线程等待队列。等待队列中会保存等待线程以及相关数据的信息。比如，生产者将数据放入SynchronousQueue时，如果没有消费者接收，那么数据本身和线程对象都会打包在队列中等待（因为SynchronousQueue容积为0，没有数据可以正常放入）。

Transferer.transfer()函数的实现是SynchronousQueue的核心，它大体上分为三个步骤：

1. 如果等待队列为空，或者队列中节点的类型和本次操作是一致

的，那么将当前操作压入队列等待。比如，等待队列中是读线程等待，本次操作也是读，因此这两个读都需要等待。进入等待队列的线程可能会被挂起，它们会等待一个“匹配”操作。

2. 如果等待队列中的元素和本次操作是互补的（比如等待操作是读，而本次操作是写），那么就插入一个“完成”状态的节点，并且让他“匹配”到一个等待节点上。接着弹出这两个节点，并且使得对应的两个线程继续执行。
3. 如果线程发现等待队列的节点就是“完成”节点，那么帮助这个节点完成任务。其流程和步骤2是一致的。

步骤1的实现如下（代码参考JDK 7u60）：

```
01 SNode h = head;
02 if (h == null || h.mode == mode) {                                // 如果队列为空
03     if (timed && nanos <= 0) {                                     // 不进行等待
04         if (h != null && h.isCancelled())
05             casHead(h, h.next);                                   // 处理取消
06     } else
07         return null;
08 } else if (casHead(h, s = snode(s, e, h, mode))) {
09     SNode m = awaitFulfill(s, timed, nanos);                     //等待，直到匹配
10     if (m == s) {                                                 // 等待被取消
11         clean(s);
12         return null;
13     }
14     if ((h = head) != null && h.next == s)
15         casHead(h, s.next);                                       // 帮助s的
```

```

16         return (mode == REQUEST) ? m.item : s.item;
17     }
18 }

```

上述代码中，第1行SNode表示等待队列中的节点。内部封装了当前线程、next节点、匹配节点、数据内容等信息。第2行，判断当前等待队列为空，或者队列中元素的模式与本次操作相同（比如，都是读操作，那么都必须要等待）。第8行，生成一个新的节点并置于队列头部，这个节点就代表当前线程。如果入队成功，则执行第9行awaitFulfill()函数。该函数会进行自旋等待，并最终挂起当前线程。直到一个与之对应的操作产生，将其唤醒。线程被唤醒后（表示已经读取到数据或者自己产生的数据已经被别的线程读取），在第14~15行尝试帮助对应的线程完成两个头部节点的出队操作（这仅仅是友情帮助）。并在最后，返回读取或者写入的数据（第16行）。

步骤2的实现如下：

```

01 } else if (!isFulfilling(h.mode)) {           //是否处于fulfill状态
02     if (h.isCancelled())                       // 如果以前取消了
03         casHead(h, h.next);                   // 弹出并重试
04     else if (casHead(h, s=snode(s, e, h, FULFILLING|mode))) {
05         for (;;) {                             // 一直循环直到匹配 (
06             SNode m = s.next;                 // m 是 s的匹配者 (m
07             if (m == null) {                   // 已经没有等待者了
08                 casHead(s, null);              // 弹出fulfill节点
09                 s = null;                      // 下一次使用新的节点
10                 break;                         // 重新开始主循环
11             }

```

```

12         SNode mn = m.next;
13         if (m.tryMatch(s)) {
14             casHead(s, mn);                // 弹出s 和 m
15             return (mode == REQUEST) ? m.item : s.item;
16         } else                            // match 失败
17             s.casNext(m, mn);              // 帮助删除节点
18     }
19 }
20 }

```

上述代码中，首先判断头部节点是否处于fulfill模式。如果是，则需要进入步骤3。否则，将视自己为对应的fulfill线程。第4行，生成一个SNode元素，设置为fulfill模式并将其压入队列头部。接着，设置m（原始的队列头部）为s的匹配节点（第13行），这个tryMatch()操作将会激活一个等待线程，并将m传递给那个线程。如果设置成功，则表示数据投递完成，将s和m两个节点弹出即可（第14行）。如果tryMatch()失败，则表示已经有其他线程帮我完成了操作，那么简单得删除m节点即可（第17行），因为这个节点的数据已经被投递，不需要再次处理，然后，再次跳转到第5行的循环体，进行下一个等待线程的匹配和数据投递，直到队列中没有等待线程为止。

步骤3的实现（如果线程在执行时，发现头部元素恰好是fulfill模式，它就会帮助这个fulfill节点尽快被执行）：

```

} else {                                // 帮助一个 fulfiller
    SNode m = h.next;                   // m 是 h的 match
    if (m == null)                       // 没有等待者
        casHead(h, null);               // 弹出fulfill节点
}

```

```

else {
    SNode mn = m.next;
    if (m.tryMatch(h))                // 尝试 match
        casHead(h, mn);                // 弹出 h 和 m
    else                               // match失败
        h.casNext(m, mn);              // 帮助删除节点
}
}

```

上述代码的执行原理和步骤2是完全一致的。唯一的不同是步骤3不会返回，因为步骤3所进行的工作是帮助其他线程尽快投递它们的数据，而自己并没有完成对应的操作。因此，线程进入步骤3后，再次进入大循环体（代码中没有给出），从步骤1开始重新判断条件和投递数据。

从整个数据投递的过程中可以看到，在SynchronousQueue中，参与工作的所有线程不仅仅是竞争资源的关系。更重要的是，它们彼此之间还会互相帮助。在一个线程内部，可能会帮助其他线程完成它们的工作。这种模式可以更大程度上减少饥饿的可能，提高系统整体的并行度。

4.5 有关死锁的问题

在学习了无锁之后，让我们重新回到锁的世界吧！在众多的应用程序中，使用锁的情况一般要多于无锁。因为对于应用来说，如果业务逻辑很复杂，会极大增加无锁的编程难度。但如果使用锁，我们就不得不面对一个新的问题引起重视——那就是死锁。

那什么是死锁呢？通俗的说，死锁就是两个或者多个线程，相互占用对方需要的资源，而都不进行释放，导致彼此之间都相互等待对方释放资源，产生了无限制等待的现象。死锁一旦发生，如果没有外力介入，这种等待将永远存在，从而对程序产生严重的影响。

用来描述死锁问题的一个有名的场景是“哲学家就餐”问题。哲学家就餐问题可以这样表述，假设有五位哲学家围坐在一张圆形餐桌旁，做以下两件事情之一：吃饭，或者思考。吃东西的时候，他们就停止思考，思考的时候也停止吃东西。餐桌中间有一大碗意大利面，每两个哲学家之间有一只餐叉。因为用一只餐叉很难吃到意大利面，所以假设哲学家必须用两只餐叉吃东西。他们只能使用自己左右手边的那两只餐叉。哲学家就餐问题有时也用米饭和筷子而不是意大利面和餐叉来描述，因为很明显，吃米饭必须用两根筷子。

哲学家从来不交谈，这就很危险，可能产生死锁，每个哲学家都拿着左手的餐叉，永远都在等右边的餐叉（或者相反）。如图4.3所示，显示了这种情况。



图4-3 哲学家就餐问题

最简单的情况就是只有两个哲学家，假设是A和B。桌面也只有两个叉子。A左手拿着其中一只叉子，B也一样。这样他们的右手等在等待对方的叉子，并且这种等待会一直持续，从而导致程序永远无法正常执行。

下面让我们用一个简单的例子来模拟这个过程：

```
01 public class DeadLock extends Thread {
02     protected Object tool;
03     static Object fork1 = new Object();
04     static Object fork2 = new Object();
05
06     public DeadLock(Object obj) {
07         this.tool = obj;
08         if (tool == fork1) {
09             this.setName("哲学家A");
10         }
```

```
11         if (tool == fork2) {
12             this.setName("哲学家B");
13         }
14     }
15
16     @Override
17     public void run() {
18         if (tool == fork1) {
19             synchronized (fork1) {
20                 try {
21                     Thread.sleep(500);
22                 } catch (Exception e) {
23                     e.printStackTrace();
24                 }
25                 synchronized (fork2) {
26                     System.out.println("哲学家A开始吃饭了");
27                 }
28             }
29
30         }
31         if (tool == fork2) {
32             synchronized (fork2) {
33                 try {
34                     Thread.sleep(500);
35                 } catch (Exception e) {
36                     e.printStackTrace();
37                 }
38             }
39         }
40     }
41 }
```

```

38         synchronized (fork1) {
39             System.out.println("哲学家B开始吃饭了");
40         }
41     }
42
43 }
44
45
46 public static void main(String[] args) throws InterruptedException
47     {
48         DeadLock 哲学家A = new DeadLock(fork1);
49         DeadLock 哲学家B = new DeadLock(fork2);
50         哲学家A.start();
51         哲学家B.start();
52         Thread.sleep(1000);
53     }

```

上述代码模拟了两个哲学家互相等待对方的叉子。哲学家A先占用叉子1，哲学家B占用叉子2，接着他们就相互等待，都没有办法同时获得两个叉子用餐。

如果在实际环境中，遇到了这种情况，通常的表现就是相关的进程不再工作，并且CPU占用率为0（因为死锁的线程不占用CPU），不过这种表面现象只能用来猜测问题。如果想要确认问题，还需要使用JDK提供的一套专业工具。

首先，我们可以使用jps命令得到java进程的进程ID，接着使用jstack命令得到线程的线程堆栈：

```
C:\Users\Administrator>jps
```

```
8404
```

```
944
```

```
3992 DeadLock
```

```
3260 Jps
```

```
//使用jstack查看进程内所有的线程堆栈
```

```
C:\Users\Administrator>jstack 3992
```

```
//省略部分输出，只列出当前与死锁有关的线程
```

```
"哲学家B" #9 prio=5 os_prio=0 tid=0x01ccf400 nid=0xb70 waiting for
```

```
java.lang.Thread.State: BLOCKED (on object monitor)
```

```
at geym.conc.ch4.deadlock.DeadLock.run(DeadLock.java:42)
```

```
- waiting to lock <0x046b3430> (a java.lang.Object)
```

```
- locked <0x046b3438> (a java.lang.Object)
```

```
"哲学家A" #8 prio=5 os_prio=0 tid=0x01ccec00 nid=0x1064 waiting fo
```

```
java.lang.Thread.State: BLOCKED (on object monitor)
```

```
at geym.conc.ch4.deadlock.DeadLock.run(DeadLock.java:29)
```

```
- waiting to lock <0x046b3438> (a java.lang.Object)
```

```
- locked <0x046b3430> (a java.lang.Object)
```

```
//自动找到了一个死锁，确认死锁的存在
```

```
Found one Java-level deadlock:
```

```
=====
```

```
"哲学家B":
```

```
waiting to lock monitor 0x15b5bd6c (object 0x046b3430, a java.l
```

```
which is held by "哲学家A"
```

```
"哲学家A":
```

```
waiting to lock monitor 0x01c1705c (object 0x046b3438, a java.l
```

```
which is held by "哲学???B"
```

```
Java stack information for the threads listed above:
```

```
=====
```

```
//哲学家A占用了0x046b3430，等待0x046b3438，哲学家B正好相反，因此产生死锁  
"哲学家B":
```

```
    at geym.conc.ch4.deadlock.DeadLock.run(DeadLock.java:42)  
    - waiting to lock <0x046b3430> (a java.lang.Object)  
    - locked <0x046b3438> (a java.lang.Object)
```

```
"哲学家A":
```

```
    at geym.conc.ch4.deadlock.DeadLock.run(DeadLock.java:29)  
    - waiting to lock <0x046b3438> (a java.lang.Object)  
    - locked <0x046b3430> (a java.lang.Object)
```

```
Found 1 deadlock.
```

上面显示了jstack的部分输出。可以看到，哲学家A和哲学家B两个线程发生了死锁。并且在最后，可以看到两者相互等待的锁的ID。同时，死锁的两个线程均处于BLOCK状态。

如果想避免死锁，除了使用无锁的函数外，另外一种有效的做法是使用第三章介绍的重入锁，通过重入锁的中断或者限时等待可以有效规避死锁带来的问题。大家可以再回顾一下相关内容。

4.6 参考文献

- 有关偏向锁、轻量级锁、自旋锁等虚拟机中的锁优化
- 有关强引用、软引用、弱引用的概念
- 有关Bootstrap ClassLoader
 - 《实战Java虚拟机——JVM故障诊断与性能优化》
- 有关Hibernate中ThreadLocal的使用
 - http://blog.sina.com.cn/s/blog_7ffb8dd5010146i3.html
- 有关CAS的指令集，可以参考
 - <http://web.itu.edu.tr/kesgin/mul06/intel/instr/cmpxchg.html>
- 有关Unsafe的使用
 - <http://www.uuencode.net/201407/java-unsafe>
- java.util.Vector的空间扩展
 - <http://www.uuencode.net/201504/vector-size-alloc>
- 有关哲学家就餐问题
 - <http://zh.wikipedia.org/wiki/哲学家就餐问题>

第5章 并行模式与算法

由于并行程序设计比串行程序复杂得多。因此，我强烈建议大家可以从熟悉和了解一些常见的设计方法。就好像练习武术一样，一招一式都是要经过学习的。如果自己胡乱打一气，效果不见得好。前人总结一些武术套路，对于初学者来说，不需要发挥自己的想象力，只要按照武术套路出拳就可以了。等到练到了一定的高度，就可以以无招胜有招了，而不必拘泥于套路。这些武术套路和招数，对应到软件开发中来，就是设计模式。在这一章中，我将重点向大家介绍一些有关并行的设计模式以及算法。这些都是前人的经验总结 and 智慧的结晶。大家可以在熟知其思想和原理的基础之上，再根据自己的需求进行扩展，可能会达到更好的效果。

5.1 探讨单例模式

单例模式是设计模式中使用最为普遍的模式之一。它是一种对象创建模式，用于产生一个对象的具体实例，它可以确保系统中一个类只产生一个实例。在Java中，这样的行为能带来两大好处：

- 对于频繁使用的对象，可以省略new操作花费的时间，这对于那些重量级对象而言，是非常可观的一笔系统开销；
- 由于new操作的次数减少，因而对系统内存的使用频率也会降低，这将减轻GC压力，缩短GC停顿时间。

严格来说，单例模式与并行没有直接的关系。这里我希望讨论这个模式，是因为它实在是太常见了。并且，我们不可避免的，会在多线程环境中使用它们。并且，系统中使用单例的地方可能非常频繁，因此，我们非常迫切需要一种高效的单例实现。

下面给出了一个单例的实现，这个实现是非常简单的，但无疑是一个正确并且良好的实现。

```
1 public class Singleton {
2     private Singleton(){
3         System.out.println("Singleton is create");
4     }
5     private static Singleton instance = new Singleton();
6     public static Singleton getInstance() {
7         return instance;
```



```
8     }  
9 }
```

使用以上方式创建单例有几点必须特别注意。因为我们要保证系统中不会有人意外创建多余的实例，因此，我们把Singleton的构造函数设置为private。这点非常重要，这就警告所有的开发人员，不能随便创建这个类的实例，从而有效避免该类被错误的创建。

第二点，instance对象必须是private并且static的。如果不是private，那么instance的安全性无法得到保证。一个小小的意外就可能使得instance变成null。其次，因为工厂方法getInstance()必须是static的，因此对应的instance也必须是static。

这个单例的性能是非常好的，因为getInstance()方法只是简单地返回instance，并没有任何锁操作，因此它在并行程序中，会有良好的表现。

但是这种方式有一点明显不足，就是Singleton构造函数，或者说Singleton实例在什么时候创建是不受控制的。对于静态成员instance，它会在类第一次初始化的时候被创建。这个时刻并不一定是getInstance()方法第一次被调用的时候。

比如，如果你的单例像是这样的：

```
public class Singleton {  
    public static int STATUS=1;  
    private Singleton(){  
        System.out.println("Singleton is create");  
    }  
}
```

```
private static Singleton instance = new Singleton();  
public static Singleton getInstance() {  
    return instance;  
}  
}
```

注意，这个单例还包含一个表示状态的静态成员STATUS。此时，在相同任何地方引用这个STATUS都会导致instance实例被创建（任何对Singleton方法或者字段的引用，都会导致类初始化，并创建instance实例，但是类初始化只有一次，因此instance实例永远只会被创建一次）。比如：

```
System.out.println(Singleton.STATUS);
```

上述println会打印出：

```
Singleton is create  
1
```

可以看到，即使系统没有要求创建单例，new Singleton()也会被调用。

如果大家觉得这个小小的不足并不重要，我认为这种单例模式是一种不错的选择。它容易实现，代码易读而且性能优越。

但如果你想精确控制instance的创建时间，那么这种方式就不太友善了。我们需要寻找一种新的方法，一种支持延迟加载的策略，它只会在instance被第一次使用时，创建对象。具体实现如下：

```
01 public class LazySingleton {
```

```
02     private LazySingleton() {
03         System.out.println("LazySingleton is create");
04     }
05     private static LazySingleton instance = null;
06     public static synchronized LazySingleton getInstance() {
07         if (instance == null)
08             instance = new LazySingleton();
09         return instance;
10     }
11 }
```

这个LazySingleton的核心思想如下：最初，我们并不需要实例化instance，而当getInstance()方法被第一次调用时，创建单例对象。为了防止对象被多次创建，我们不得不使用synchronized进行方法同步。这种实现的好处是，充分利用了延迟加载，只在真正需要时创建对象。但坏处也很明显，并发环境下加锁，竞争激烈的场合对性能可能产生一定的影响。但总体上，这是一个非常易于实现和理解的方法。

此外，还有一种被称为双重检查模式的方法可以用于创建单例。但我并不打算在这里介绍它，因为这是一种非常丑陋、复杂的方法，甚至在低版本的JDK中都不能保证正确性。因此，绝不推荐大家使用。如果大家阅读到相关文档，我也强烈建议大家不要在这种方法上花费太多时间。

在上述介绍的两种单例实现中，可以说是各有千秋。有没有一种方法可以结合二者之优势呢？答案是肯定的：

```
01 public class StaticSingleton {
```

```
02     private StaticSingleton(){
03         System.out.println("StaticSingleton is create");
04     }
05     private static class SingletonHolder {
06         private static StaticSingleton instance = new StaticSi
07     }
08     public static StaticSingleton getInstance() {
09         return SingletonHolder.instance;
10     }
11 }
```

上述代码实现了一个单例，并且同时拥有前两种方式的有点。首先 `getInstance()` 方法中没有锁，这使得在高并发环境下性能优越。其次，只有在 `getInstance()` 方法被第一次调用时，`StaticSingleton` 的实例才会被创建。因为这种方法巧妙地使用了内部类和类的初始化方式。内部类 `SingletonHolder` 被申明为 `private`，这使得我们不可能在外部访问并初始化它。而我们只可能在 `getInstance()` 内部对 `SingletonHolder` 类进行初始化，利用虚拟机的类初始化机制创建单例。

5.2 不变模式

在并行软件开发过程中，同步操作似乎是必不可少的。当多线程对同一个对象进行读写操作时，为了保证对象数据的一致性和正确性，有必要对对象进行同步。而同步操作对系统性能是有相当的损耗。为了尽可能地去除这些同步操作，提高并行程序性能，可以使用一种不可改变的对象，依靠对象的不变性，可以确保其在没有同步操作的多线程环境中依然始终保持内部状态的一致性和正确性。这就是不变模式。

不变模式天生就是多线程友好的，它的核心思想是，一个对象一旦被创建，则它的内部状态将永远不会发生改变。所以，没有一个线程可以修改其内部状态和数据，同时其内部状态也绝不会自行发生改变。基于这些特性，对不变对象的多线程操作不需要进行同步控制。

同时还需要注意，不变模式和只读属性是有一定的区别的。不变模式是比只读属性具有更强的一致性和不变性。对只读属性的对象而言，对象本身不能被其他线程修改，但是对象的自身状态却可能自行修改。

比如，一个对象的存活时间（对象创建时间和当前时间的时间差）是只读的，因为任何一个第三方线程都不能修改这个属性，但是这是一个可变的属性，因为随着时间的推移，存活时间时刻都在发生变化。而不变模式则要求，无论出于什么原因，对象自创建后，其内部状态和数据保持绝对的稳定。

因此，不变模式的主要使用场景需要满足以下2个条件：

- 当对象创建后，其内部状态和数据不再发生任何变化。

- 对象需要被共享，被多线程频繁访问。

在Java语言中，不变模式的实现很简单。为确保对象被创建后，不发生任何改变，并保证不变模式正常工作，只需要注意以下4点：

- 去除setter方法以及所有修改自身属性的方法。
- 将所有属性设置为私有，并用final标记，确保其不可修改。
- 确保没有子类可以重载修改它的行为。
- 有一个可以创建完整对象的构造函数。

以下代码实现了一个不变的产品对象，它拥有序列号、名称和价格三个属性。

```
public final class Product {                                //确保无子类
    private final String no;                                //私有属性,
    private final String name;                              //final保
    private final double price;

    public Product(String no, String name, double price) { //在
        super();                                           //因为创建
        this.no = no;
        this.name = name;
        this.price = price;
    }

    public String getNo() {
        return no;
    }
}
```

```
public String getName() {  
    return name;  
}  
public double getPrice() {  
    return price;  
}  
}
```

在不变模式的实现中，**final**关键字起到了重要的作用。对属性的**final**定义确保所有数据只能在对象被构造时赋值1次。之后，就永远不再发生改变。而对**class**的**final**确保了类不会有子类。根据里氏代换原则，子类可以完全的替代父类。如果父类是不变的，那么子类也必须是不变的，但实际上我们并无法约束这点，为了防止子类做出一些意外的行为，这里就干脆把子类都禁用了。

在JDK中，不变模式的应用非常广泛。其中，最为典型的就是 `java.lang.String` 类。此外，所有的元数据类包装类，都是使用不变模式实现的。主要的不变模式类型如下：

- `java.lang.String`
- `java.lang.Boolean`
- `java.lang.Byte`
- `java.lang.Character`
- `java.lang.Double`
- `java.lang.Float`
- `java.lang.Integer`
- `java.lang.Long`
- `java.lang.Short`

由于基本数据类型和String类型在实际的软件开发中应用极其广泛，使用不变模式后，所有实例的方法均不需要进行同步操作，保证了它们在多线程环境下的性能。

注意：不变模式通过回避问题而不是解决问题的态度来处理多线程并发访问控制。不变对象是不需要进行同步操作的。由于并发同步会对性能产生不良的影响，因此，在需求允许的情况下，不变模式可以提高系统的并发性能和并发量。

5.3 生产者-消费者模式

生产者-消费者模式是一个经典的多线程设计模式，它为多线程间的协作提供了良好的解决方案。在生产者-消费者模式中，通常有两类线程，即若干个生产者线程和若干个消费者线程。生产者线程负责提交用户请求，消费者线程则负责具体处理生产者提交的任务。生产者和消费者之间则通过共享内存缓冲区进行通信。

如图5.1所示，展示了生产者-消费者模式的基本结构。三个生产者线程将任务提交到共享内存缓冲区，消费者线程并不直接与生产者线程通信，而在共享内存缓冲区中获取任务，并进行处理。

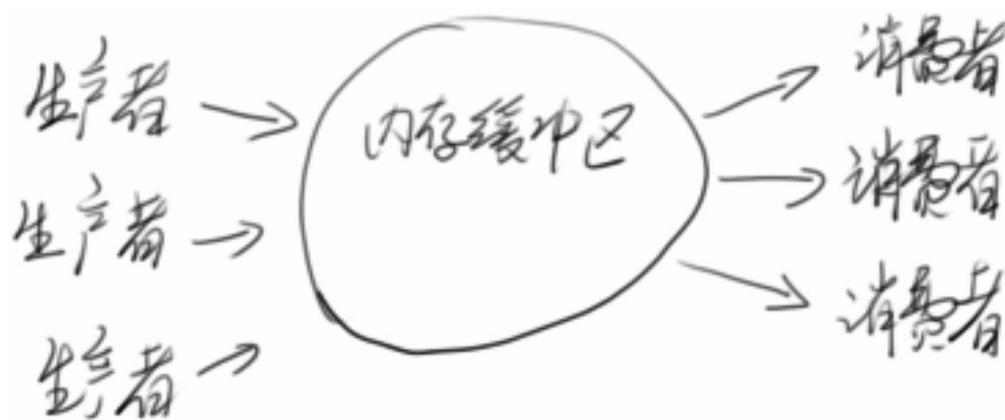


图5.1 生产者-消费者模式架构图

注意：生产者-消费者模式中的内存缓存区的主要功能是数据在多线程间的共享，此外，通过该缓冲区，可以缓解生产者和消费者间的性能差。

生产者-消费者模式的核心组件是共享内存缓存区，它作为生产者和消费者间的通信桥梁，避免了生产者和消费者的直接通信，从而将生

产者和消费者进行解耦。生产者不需要知道消费者的存在，消费者也不需要知道生产者的存在。

同时，由于内存缓冲区的存在，允许生产者和消费者在执行速度上存在时间差，无论是生产者在某一局部时间内速度高于消费者，还是消费者在局部时间内高于生产者，都可以通过共享内存缓冲区得到缓解，确保系统正常运行。

生产者-消费者模式的主要角色如表5.1所示。

表5.1 生产者-消费者模式主要角色

角色	作用
生产者	用于提交用户请求，提取用户任务，并装入内存缓冲区
消费者	在内存缓冲区中提取并处理任务
内存缓冲区	缓存生产者提交的任务或数据，供消费者使用
任务	生产者向内存缓冲区提交的数据结构
Main	使用生产者和消费者的客户端

图5.2显示了生产者-消费者模式一种实现的具体结构。

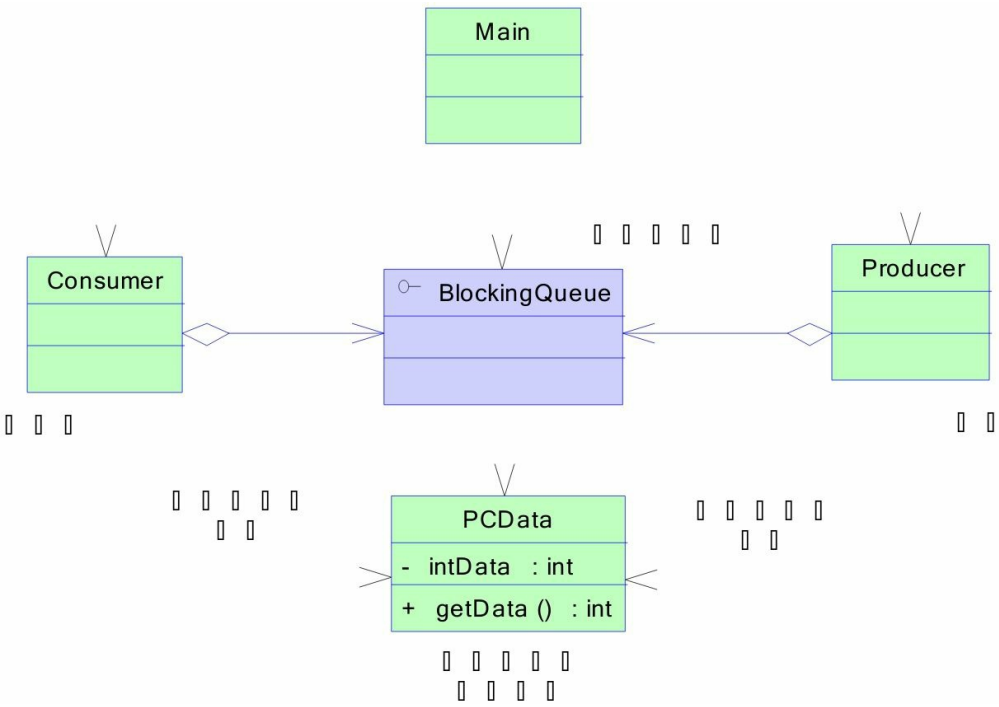


图5.2 生产者-消费者实现类图

其中，BlockingQueue充当了共享内存缓冲区，用于维护任务或数据队列（PCData对象）。我强烈建议大家先回顾一下第3章有关BlockingQueue的相关知识，对于理解整个生产者和消费者结构有重要的帮助。PCData对象表示一个生产任务，或者相关任务的数据。生产者对象和消费者对象均引用同一个BlockingQueue实例。生产者负责创建PCData对象，并将它加入BlockingQueue中，消费者则从BlockingQueue队列中获取PCData。

基于图5.2所示结构，实现一个基于生产者-消费者模式的求整数平方的并程序。

首先，生产者线程的实现如下，它构建PCData对象，并放入BlockingQueue队列中。

```
public class Producer implements Runnable {
    private volatile boolean isRunning = true;
    private BlockingQueue<PCData> queue;
    private static AtomicInteger count = new AtomicInteger();
    private static final int SLEEPTIME = 1000;

    public Producer(BlockingQueue<PCData> queue) {
        this.queue = queue;
    }

    public void run() {
        PCData data = null;
```

```

    Random r = new Random();

    System.out.println("start producer id="+Thread.currentThread().getId());
    try {
        while (isRunning) {
            Thread.sleep(r.nextInt(SLEEPTIME));
            data = new PCData(count.incrementAndGet());
            System.out.println(data+" is put into queue");
            if (!queue.offer(data, 2, TimeUnit.SECONDS)) {
                System.err.println("failed to put data: " + data);
            }
        }
    } catch (InterruptedException e) {
        e.printStackTrace();
        Thread.currentThread().interrupt();
    }
}

public void stop() {
    isRunning = false;
}
}

```

对应的消费者的实现如下。它从BlockingQueue队列中取出PCData对象，并进行相应的计算。

```

public class Consumer implements Runnable {
    private BlockingQueue<PCData> queue;
    //...
}

```

```

private static final int SLEEPTIME = 1000;

public Consumer(BlockingQueue<PCData> queue) {
    this.queue = queue;
}

public void run() {
    System.out.println("start Consumer id="
        + Thread.currentThread().getId());
    Random r = new Random(); //随机等

    try {
        while(true){
            PCData data = queue.take(); //提取任
            if (null != data) {
                int re = data.getData() * data.getData();
                System.out.println(MessageFormat.format("{0}*"
                    data.getData(), data.getData(), re));
                Thread.sleep(r.nextInt(SLEEPTIME));
            }
        }
    } catch (InterruptedException e) {
        e.printStackTrace();
        Thread.currentThread().interrupt();
    }
}
}

```

PCData作为生产者和消费者之间的共享数据模型，定义如下：

```
public final class PCData {                                //任务相关
    private final int intData;                             //数据
    public PCData(int d){
        intData=d;
    }
    public PCData(String d){
        intData=Integer.valueOf(d);
    }
    public int getData(){
        return intData;
    }
    @Override
    public String toString(){
        return "data:"+intData;
    }
}
```

在主函数中，创建三个生产者和三个消费者，并让它们协作运行。在主函数的实现中，定义LinkedBlockingQueue作为BlockingQueue的实现类。

```
public class Main {
    public static void main(String[] args) throws InterruptedException
        //建立缓冲区
        BlockingQueue<PCData> queue = new LinkedBlockingQueue<P
```

```
    Producer producer1 = new Producer(queue);
    Producer producer2 = new Producer(queue);
    Producer producer3 = new Producer(queue);
    Consumer consumer1 = new Consumer(queue);
    Consumer consumer2 = new Consumer(queue);
    Consumer consumer3 = new Consumer(queue);
    ExecutorService service = Executors.newCachedThreadPool()
    service.execute(producer1);
    service.execute(producer2);
    service.execute(producer3);
    service.execute(consumer1);
    service.execute(consumer2);
    service.execute(consumer3);
    Thread.sleep(10 * 1000);
    producer1.stop();
    producer2.stop();
    producer3.stop();
    Thread.sleep(3000);
    service.shutdown();
}
}
```

注意：生产者-消费者模式很好地对生产者线程和消费者线程进行解耦，优化了系统整体结构。同时，由于缓冲区的作用，允许生产者线程和消费者线程存在执行上的性能差异，从一定程度上缓解了性能瓶颈对系统性能的影响。

5.4 高性能的生产者-消费者：无锁的实现

BlockigQueue用于实现生产者和消费者一个不错的选择。它可以很自然地实现作为生产者和消费者的内存缓冲区。但是BlockigQueue并不是一个高性能的实现，它完全使用锁和阻塞等待来实现线程间的同步。在高并发场合，它的性能并不是特别的优越。就像之前我已经提过的：ConcurrentLinkedQueue是一个高性能的队列，但是BlockingQueue只是为了方便数据共享。

而ConcurrentLinkedQueue的秘诀就在于大量使用了无锁的CAS操作。同理，如果我们使用CAS来实现生产者-消费者模式，也同样可以获得可观的性能提升。不过正如大家所见，使用CAS进行编程是非常困难的，但有一个好消息是，目前有一个现成的Disruptor框架，它已经帮助我们实现了这一个功能。

5.4.1 无锁的缓存框架：Disruptor

Disruptor框架是由LMAX公司开发的一款高效的无锁内存队列。它使用无锁的方式实现了一个环形队列，非常适合于实现生产者和消费者模式，比如事件和消息的发布。在Disruptor中，别出心裁地使用了环形队列（RingBuffer）来代替普通线性队列，这个环形队列内部实现为一个普通的数组。对于一般的队列，势必要提供队列同步head和尾部tail两个指针，用于出队和入队，这样无疑就增加了线程协作的复杂度。但如

果队列是环形的，则只需要对外提供一个当前位置`cursor`，利用这个指针既可以进入入队也可以进行出队操作。由于环形队列的缘故，队列的总大小必须事先指定，不能动态扩展。为了能够快速从一个序列（`sequence`）对应到数组的实际位置（每次有元素入队，序列就加1），Disruptor要求我们必须将数组的大小设置为2的整数次方。这样通过 `sequence & (queueSize-1)` 就能立即定位到实际的元素位置`index`。这个要比取余（%）操作快得多。

如果大家不理解上面的`sequence & (queueSize-1)`，我在这里再简单说明一下。如果`queueSize`是2的整数次幂，则这个数字的二进制表示必然是10、100、1000、10000等形式。因此，`queueSize-1`的二进制则是一个全1的数字。因此它可以将`sequence`限定在`queueSize-1`范围内，并且不会有任何一位是浪费的。

如图5.3所示，显示了RingBuffer的结构。生产者向缓冲区中写入数据，而消费者从中读取数据。生产者写入数据时，使用CAS操作，消费者读取数据时，为了防止多个消费者处理同一个数据，也使用CAS操作进行数据保护。

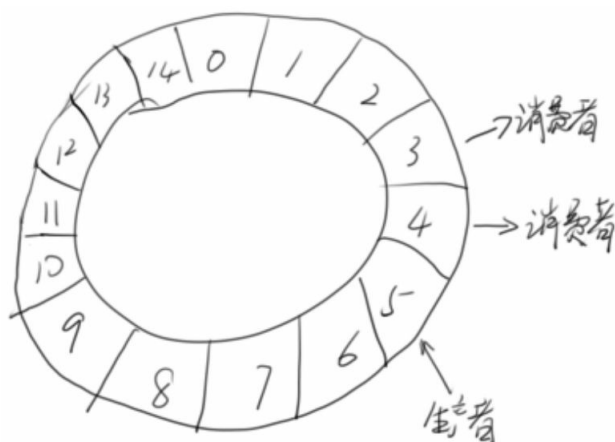


图5.3 Disruptor的RingBuffer结构

这种固定大小的环形队列的另外一个好处就是可以做到完全的内存复用。在系统的运行过程中，不会有新的空间需要分配或者老的空间需要回收。因此，可以大大减少系统分配空间以及回收空间的额外开销。

5.4.2 用Disruptor实现生产者-消费者案例

现在我们已经基本了解了Disruptor的基本实现。在本节，我们将展示一下Disruptor的基本使用和API，这里，我们使用的版本是disruptor-3.3.2，不同版本的disruptor可能会有细微的差别，也请大家留意。

这里，我们的生产者不断产生整数，消费者读取生产者的数据，并计算其平方。

首先，我们还是需要一个代表数据的PCData：

```
public class PCData
{
    private long value;
    public void set(long value)
    {
        this.value = value;
    }
    public long get(){
        return value;
    }
}
```

消费者实现为WorkHandler接口，它来自Disruptor框架：

```
public class Consumer implements WorkHandler<PCData> {  
    @Override  
    public void onEvent(PCData event) throws Exception {  
        System.out.println(Thread.currentThread().getId() + ":Eve  
            + event.get() * event.get() + "--");  
    }  
}
```

消费者的作用是读取数据进行处理。这里，数据的读取已经由Disruptor进行封装，onEvent()方法为框架的回调方法。因此，这里只需要简单地进行数据处理即可。

还需要一个产生PCData的工厂类。它会在Disruptor系统初始化时，构造所有的缓冲区中的对象实例（之前说过Disruptor会预先分配空间）：

```
public class PCDataFactory implements EventFactory<PCData>  
{  
    public PCData newInstance()  
    {  
        return new PCData();  
    }  
}
```

接着，让我们来看一下生产者，它比前面几个类稍微复杂一点：

```
01 public class Producer
```

```

02 {
03     private final RingBuffer<PCData> ringBuffer;
04
05     public Producer(RingBuffer<PCData> ringBuffer)
06     {
07         this.ringBuffer = ringBuffer;
08     }
09
10     public void pushData(ByteBuffer bb)
11     {
12         long sequence = ringBuffer.next(); // Grab the next s
13         try
14         {
15             PCData event = ringBuffer.get(sequence); // Get th
16                                                         // for
17             event.set(bb.getLong(0)); // Fill with data
18         }
19         finally
20         {
21             ringBuffer.publish(sequence);
22         }
23     }
24 }

```

生产者需要一个RingBuffer的引用，也就是环形缓冲区。它有一个重要的方法pushData()将产生的数据推入缓冲区。方法pushData()接收一个ByteBuffer对象。在ByteBuffer中可以用来包装任何数据类型。这里用

来存储long整数，pushData()的功能就是将传入的ByteBuffer中的数据提取出来，并装载到环形缓冲区中。

上述第12行代码，通过next()方法得到下一个可用的序列号。通过序列号，取得下一个空闲可用的PCData，并且将PCData的数据设为期望值，这个值最终会传递给消费者。最后，在第21行，进行数据发布。只有发布后的数据才会真正被消费者看见。

至此，我们的生产者、消费者和数据都已经准备就绪。只差一个统筹规划的主函数将所有内容整合起来：

```
01 public static void main(String[] args) throws Exception
02 {
03     Executor executor = Executors.newCachedThreadPool();
04     PCDataFactory factory = new PCDataFactory();
05     // Specify the size of the ring buffer, must be power of 2
06     int bufferSize = 1024;
07     Disruptor<PCData> disruptor = new Disruptor<PCData>(factory,
08         bufferSize,
09         executor,
10         ProducerType.MULTI,
11         new BlockingWaitStrategy()
12     );
13     disruptor.handleEventsWithWorkerPool(
14         new Consumer(),
15         new Consumer(),
16         new Consumer(),
17         new Consumer());
```

```

18     disruptor.start();
19
20     RingBuffer<PCData> ringBuffer = disruptor.getRingBuffer()
21     Producer producer = new Producer(ringBuffer);
22     ByteBuffer bb = ByteBuffer.allocate(8);
23     for (long l = 0; true; l++)
24     {
25         bb.putLong(0, l);
26         producer.pushData(bb);
27         Thread.sleep(100);
28         System.out.println("add data "+l);
29     }
30 }

```

上述代码第6行，设置缓冲区大小为1024。显然是2的整数次幂——一个合理的大小。第7~12创建了disruptor对象。它封装了整个disruptor库的使用，提供了一些便捷的API。第13~17行，设置了用于处理数据的消费者。这里设置了4个消费者实例，系统会为将每一个消费者实例映射到一个线程中，也就是这里提供了4个消费者线程。第18行，启动并初始化disruptor系统。在第23~29行中，由一个生产者不断地向缓冲区中存入数据。

系统执行后，你就可以得到类似以下的输出：

```

8:Event: --0--
add data 0
11:Event: --1--
add data 1

```

```
10:Event: --4--  
add data 2  
9:Event: --9--  
add data 3
```

生产者和消费者正常工作。根据Disruptor的官方报告，Disruptor的性能要比BlockingQueue至少高一个数量级以上。如此诱人的性能，当然值得我们去尝试！

5.4.3 提高消费者的响应时间：选择合适的策略

当有新数据在Disruptor的环形缓冲区中产生时，消费者如何知道这些新产生的数据呢？或者说，消费者如何监控缓冲区中的信息呢？为此，Disruptor提供了几种策略，这些策略由WaitStrategy接口进行封装，主要有以下几种实现。

- **BlockingWaitStrategy**：这是默认的策略。使用BlockingWaitStrategy和使用BlockingQueue是非常类似的，它们都使用锁和条件（Condition）进行数据的监控和线程的唤醒。因为涉及到线程的切换，BlockingWaitStrategy策略是最节省CPU，但是在高并发下性能表现最糟糕的一种等待策略。
- **SleepingWaitStrategy**：这个策略也是对CPU使用率非常保守的。它会在循环中不断等待数据。它会先进行自旋等待，如果不成功，则使用Thread.yield()让出CPU，并最终使用LockSupport.parkNanos(1)进行线程休眠，以确保不占用太多的

CPU数据。因此，这个策略对于数据处理可能产生比较高的平均延时。它比较适合于对延时要求不是特别高的场合，好处是它对生产者线程的影响最小。典型的应用场景是异步日志。

- **YieldingWaitStrategy:** 这个策略用于低延时的场合。消费者线程会不断循环监控缓冲区变化，在循环内部，它会使用 `Thread.yield()` 让出CPU给别的线程执行时间。如果你需要一个高性能的系统，并且对延时有较为严格的要求，则可以考虑这种策略。使用这种策略时，相当于你的消费者线程变身成为了一个内部执行了 `Thread.yield()` 的死循环。因此，你最好有多于消费者线程数量的逻辑CPU数量（这里的逻辑CPU，我指的是“双核四线程”中的那个四线程，否则，整个应用程序恐怕都会受到影响。
- **BusySpinWaitStrategy:** 这个是最疯狂的等待策略了。它就是一个死循环！消费者线程会尽最大努力疯狂监控缓冲区的变化。因此，它会吃掉所有的CPU资源。你只有在对延迟非常苛刻的场合可以考虑使用它（或者说，你的系统真的非常繁忙）。因为在这里你等同开启了一个死循环监控，所以，你的物理CPU数量必须要大于消费者线程数。注意，我这里说的是物理CPU，如果你在一个物理核上使用超线程技术模拟两个逻辑核，另外一个逻辑核显然会受到这种超密集计算的影响而不能正常工作。

在上面的例子中，使用的是 `BlockingWaitStrategy`（第11行）。读者可以替换这个实现，体验一下不同等待策略的效果。

5.4.4 CPU Cache的优化：解决伪共享问题

除了使用CAS和提供了各种不同的等待策略来提高系统的吞吐量外。Disruptor大有将优化进行到底的气势，它甚至尝试解决CPU缓存的伪共享问题。

什么是伪共享问题呢？我们知道，为了提高CPU的速度，CPU有一个高速缓存Cache。在高速缓存中，读写数据的最小单位为缓存行（Cache Line），它是从主存（memory）复制到缓存（Cache）的最小单位，一般为32字节到128字节。

如果两个变量存放在一个缓存行中时，在多线程访问中，可能会相互影响彼此的性能。如图5.4所示，假设X和Y在同一个缓存行。运行在CPU1上的线程更新了X，那么CPU2上的缓存行就会失效，同一行的Y即使没有修改也会变成无效，导致Cache无法命中。接着，如果在CPU2上的线程更新了Y，则导致CPU1上的缓存行又失效（此时，同一行的X又变得无法访问）。这种情况反反复复发生，无疑是一个潜在的性能杀手。如果CPU经常不能命中缓存，那么系统的吞吐量就会急剧下降。

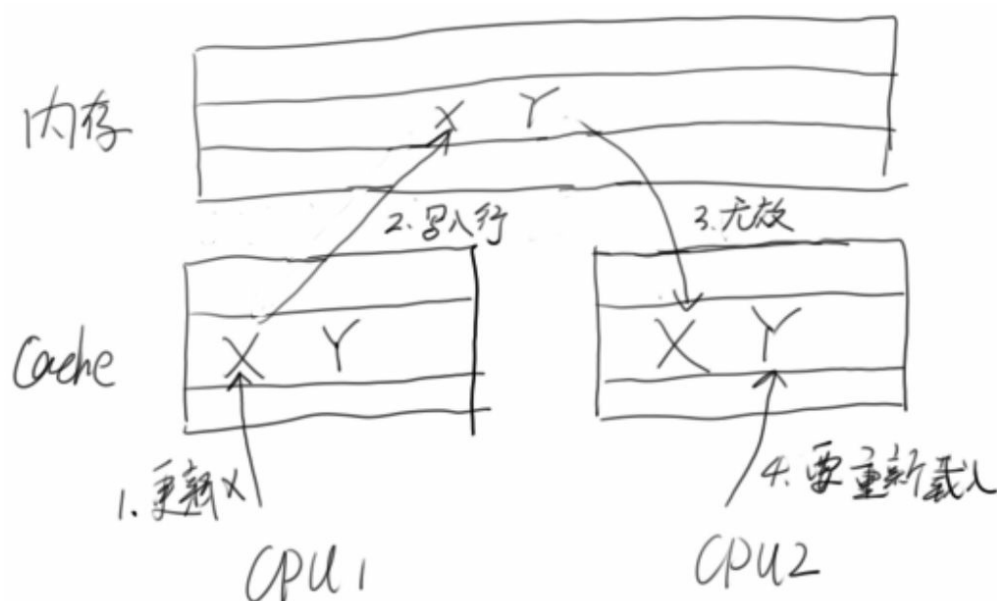


图5.4 X和Y在同一个缓存行中

为了使这种情况不发生，一种可行的做法就是在X变量的前后空间都先占据一定的位置（把它叫做padding吧，用来填充用的）。这样，当内存被读入缓存中时，这个缓存行中，只有X一个变量实际是有效的，因此就不会发生多个线程同时修改缓存行中不同变量而导致变量全体失效的情况，如图5.5所示。

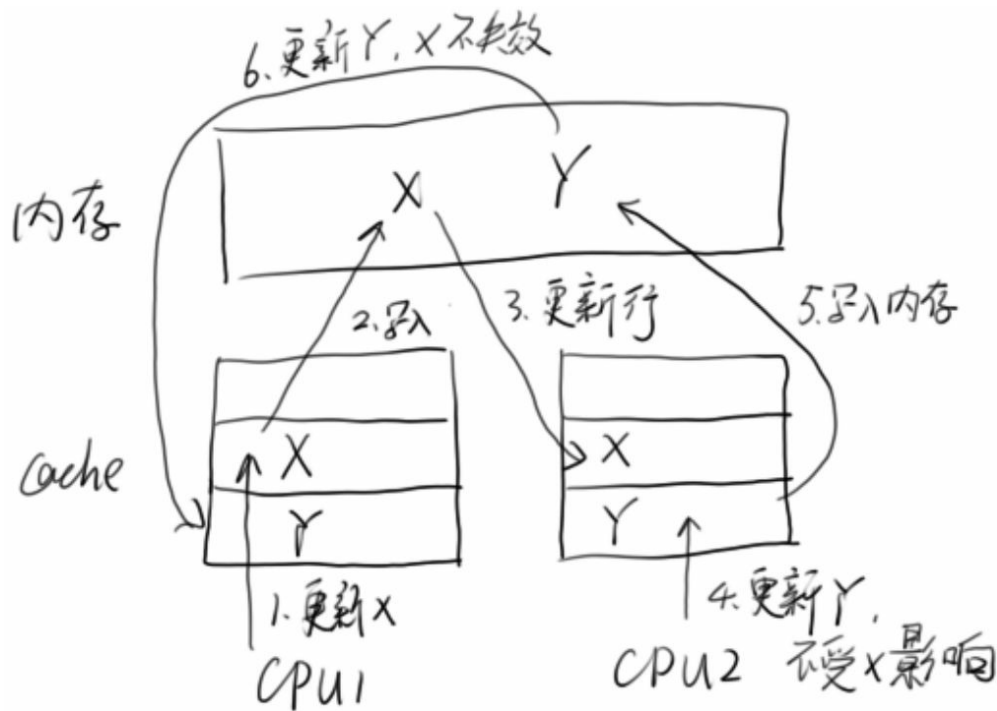


图5.5 变量X和Y各占据一个缓冲行

为了实现这个目的，我们可以这么做：

```
01 public final class FalseSharing implements Runnable {
02     public final static int NUM_THREADS = 2; // change
03     public final static long ITERATIONS = 500L * 1000L * 1000L
04     private final int arrayIndex;
05
06     private static VolatileLong[] longs = new VolatileLong[NUM
07     static {
```

```
08         for (int i = 0; i < longs.length; i++) {
09             longs[i] = new VolatileLong();
10         }
11     }
12
13     public FalseSharing(final int arrayIndex) {
14         this.arrayIndex = arrayIndex;
15     }
16
17     public static void main(final String[] args) throws Except
18         final long start = System.currentTimeMillis();
19         runTest();
20         System.out.println("duration = " + (System.currentTime
21     }
22
23     private static void runTest() throws InterruptedException
24         Thread[] threads = new Thread[NUM_THREADS];
25
26         for (int i = 0; i < threads.length; i++) {
27             threads[i] = new Thread(new FalseSharing(i));
28         }
29
30         for (Thread t : threads) {
31             t.start();
32         }
33
34         for (Thread t : threads) {
```

```

35         t.join();
36     }
37 }
38
39 public void run() {
40     long i = ITERATIONS + 1;
41     while (0 != --i) {
42         longs[arrayIndex].value = i;
43     }
44 }
45
46 public final static class VolatileLong {
47     public volatile long value = 0L;
48     public long p1, p2, p3, p4, p5, p6,p7; // comment out
49 }
50 }

```

这里我们使用两个线程，因为我的计算机是双核的，大家可以根据自己的硬件配置修改参数NUM_THREADS（第2行）。我们准备一个数组longs（第6行），数组元素个数和线程数量一致。每个线程都会访问自己对应的longs中的元素（从第42行、第27行和第14行可以看到这一点）。

最后，最关键的一点就是VolatileLong。在第48行，准备了7个long型变量用来填充缓存。实际上，只有VolatileLong.value是会被使用的。而那些p1、p2等仅仅用于将数组中第一个VolatileLong.value和第二个VolatileLong.value分开，防止它们进入同一个缓存行。

这里，我使用JDK7 64位的Java虚拟机，执行上述程序，输出如下：

```
duration = 5207
```

这说明系统花费了5秒钟完成所有的操作。如果我注释掉第48行，也就是允许系统中两个VolatileLong.value放置在同一个缓存行中，程序输出如下：

```
duration = 13675
```

很明显，第48行的填充对系统的性能是非常有帮助的。

注意：由于各个JDK版本内部实现不一致，在某些JDK版本中（比如JDK 8），会自动优化不使用的字段。这将直接导致这种padding的伪共享解决方案失效。更多详细内容大家可以参考第6章中有关LongAddr的介绍。

Disruptor框架充分考虑了这个问题，它的核心组件Sequence会被非常频繁的访问（每次入队，它都会被加1），其基本结构如下：

```
class LhsPadding
{
    protected long p1, p2, p3, p4, p5, p6, p7;
}

class Value extends LhsPadding
{
    protected volatile long value;
```

```

}

class RhsPadding extends Value
{
    protected long p9, p10, p11, p12, p13, p14, p15;
}p
ublic class Sequence extends RhsPadding{
//省略具体实现
}

```

虽然在Sequence中，主要使用的只有value。但是，通过LhsPadding和RhsPadding，在这个value的前后安置了一些占位空间，使得value可以无冲突的存在于缓存中。

此外，对于Disruptor的环形缓冲区RingBuffer，它内部的数组是通过以下语句构造的：

```
this.entries = new Object[sequencer.getBufferSize() + 2 * BUFFER_
```

大家注意，实际产生的数组大小是缓冲区实际大小再加上两倍的BUFFER_PAD。这就相当于在这个数组的头部和尾部两段各增加了BUFFER_PAD个填充，使得整个数组被载入Cache时不会受到其他变量的影响而失效。

5.5 Future模式

Future模式是多线程开发中非常常见的一种设计模式，它的核心思想是异步调用。当我们需要调用一个函数方法时，如果这个函数执行很慢，那么我们就要进行等待。但有时候，我们可能并不急着要结果。因此，我们可以让被调者立即返回，让它在后台慢慢处理这个请求。对于调用者来说，则可以先处理一些其他任务，在真正需要数据的场合再去尝试获得需要的数据。

Future模式有点类似在网上买东西。如果我们在网上下单买了一个手机，当我们支付完成后，手机并没有办法立即送到家里，但是在电脑上会立即产生一个订单。这个订单就是将来发货或者领取手机的重要凭证，这个凭证也就是Future模式中会给出的一个契约。在支付活动结束后，大家不会傻傻地等着手机到来，而是可以各忙各的。而这张订单就成为了商家配货、发货的驱动力。当然，这一切你并不用关心。你要做的，只是在快递上门时，开一下门，拿一下货而已。

对于Future模式来说，虽然它无法立即给出你需要的数据。但是，它会返回给你一个契约，将来，你可以凭借着这个契约去重新获取你需要的信息。

如图5.6所示，显示了通过传统的同步方法，调用一段比较耗时的程序。客户端发出call请求，这个请求需要相当长一段时间才能返回。客户端一直等待，直到数据返回，随后，再进行其他任务的处理。

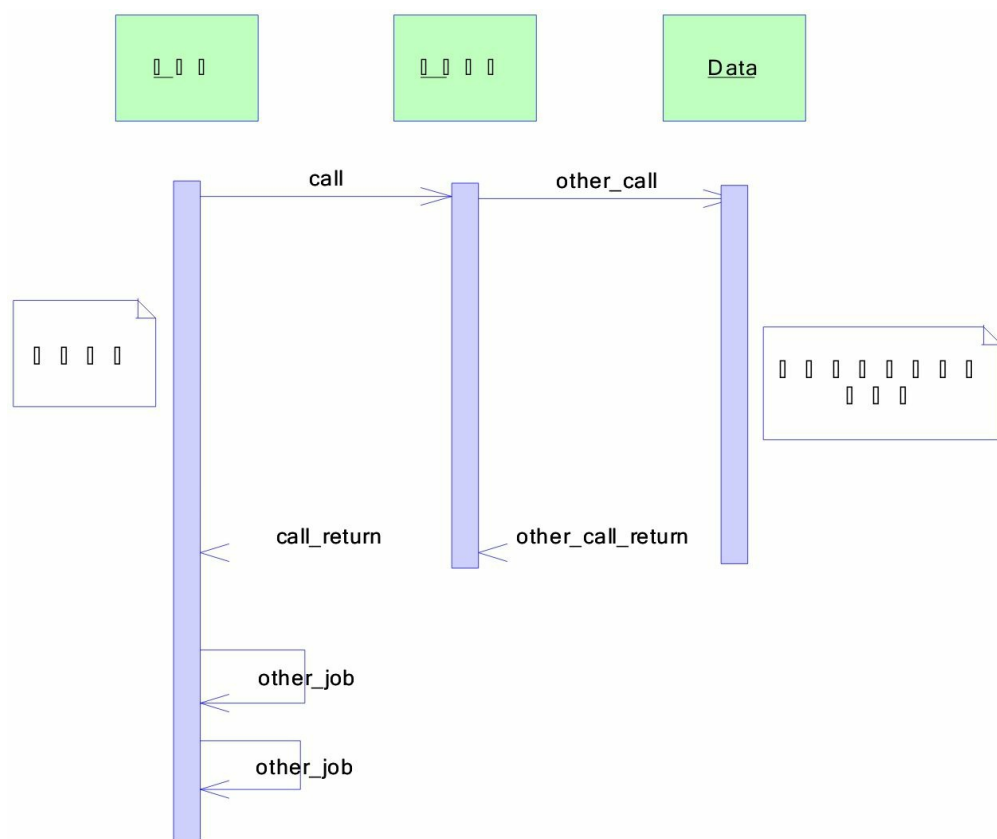


图5-6 传统串行程序调用流程

使用Future模式替换原来的实现方式，可以改进其调用过程，如图5.7所示。

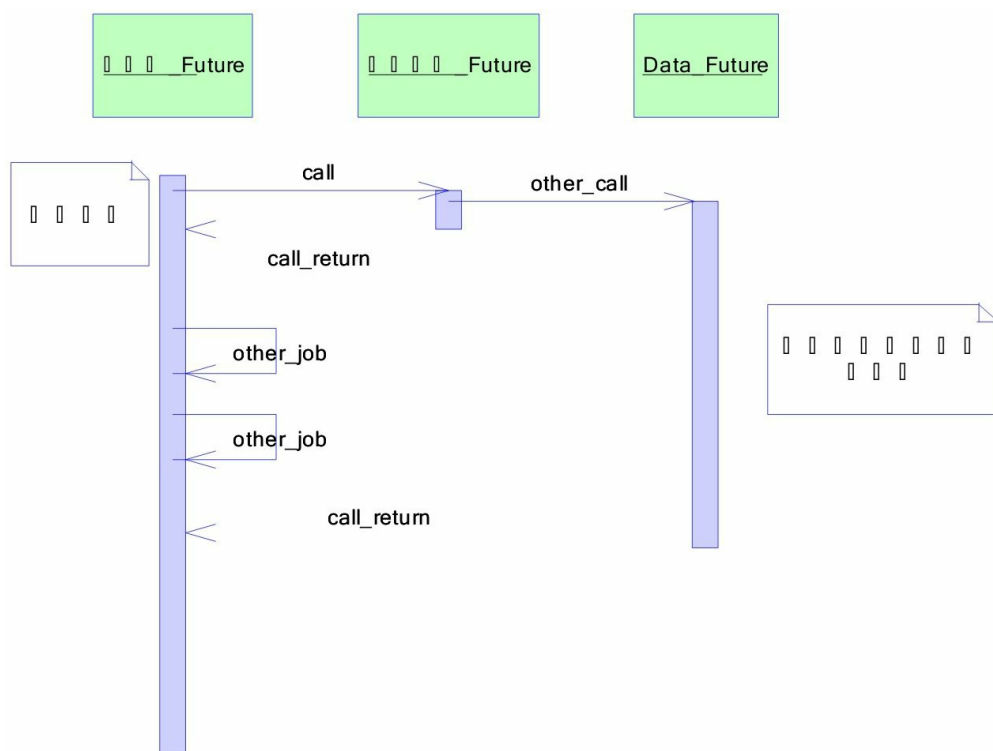


图5-7 Future模式流程图

下面的模型展示了一个广义Future模式的实现，从Data_Future对象可以看到，虽然call本身仍然需要很长一段时间处理程序。但是，服务程序不等数据处理完成便立即返回客户端一个伪造的数据（相当于商品的订单，而不是商品本身），实现了Future模式的客户端在拿到这个返回结果后，并不急于对其进行处理，而去调用了其他业务逻辑，充分利用了等待时间，这就是Future模式的核心所在。在完成了其他业务逻辑的处理后，最后再使用返回比较慢的Future数据。这样，在整个调用过程中，就不存在无谓的等待，充分利用了所有的时间片段，从而提高系统的响应速度。

5.5.1 Future模式的主要角色

为了让大家能够更清晰地认识Future模式的基本结构。在这里，我

给出一个非常简单的Future模式的实现，它的主要参与者如表5.2所示。

表5.2 Future模式的主要参与者

参与者	作用
Main	系统启动，调用Client发出请求
Client	返回Data对象，立即返回FutureData，并开启ClientThread线程装配RealData
Data	返回数据的接口
FutureData	Future数据，构造很快，但是是一个虚拟的数据，需要装配RealData
RealData	真实数据，其构造是比较慢的

它的核心结构如图5.8所示。

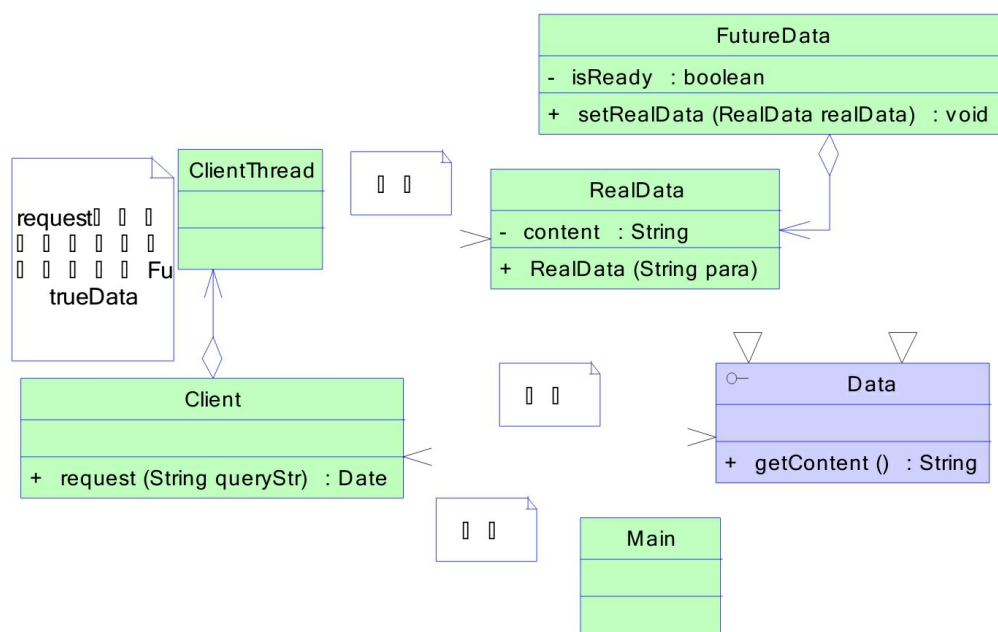


图5.8 Future模式结构图

5.5.2 Future模式的简单实现

在这个实现中，有一个核心接口Data，这就是客户端希望获取的数据。在Future模式中，这个Data接口有两个重要的实现，分别是RealData，也就是真实数据，这就是我们最终需要获得的，有价值的信息。另外一个就是FutureData，它就是用来提取RealData的一个“订单”。

因此FutureData是可以立即返回得到的。

下面是Data接口：

```
public interface Data {
    public String getResult ();
}
```

FutureData实现了一个快速返回的RealData包装。它只是一个包装，或者说是一个RealData的虚拟实现。因此，它可以很快被构造并返回。当使用FutureData的getResult()方法时，如果实际的数据没有准备好，那么程序就会阻塞，等待RealData准备好并注入到FutureData中，才最终返回数据。

注意：FutureData是Future模式的关键。它实际上是真实数据RealData的代理，封装了获取RealData的等待过程。

```
public class FutureData implements Data {
    protected RealData realdata = null;           //FutureData是F
    protected boolean isReady = false;
    public synchronized void setRealData(RealData realdata) {
        if (isReady) {
            return;
        }
        this.realdata = realdata;
        isReady = true;
        notifyAll();                               //RealData已经被
    }
}
```

```

    public synchronized String getResult() {           //会等待RealData
        while (!isReady) {
            try {
                wait();                                //一直等待，知道R
            } catch (InterruptedException e) {
            }
        }
        return realdata.result;                        //由RealData实现
    }
}

```

RealData是最终需要使用的数据模型。它的构造很慢。在这里，使用sleep()函数模拟这个过程，简单地模拟一个字符串的构造。

```

public class RealData implements Data {
    protected final String result;
    public RealData(String para) {
        //RealData的构造可能很慢，需要用户等待很久，这里使用sleep模拟
        StringBuffer sb=new StringBuffer();
        for (int i = 0; i < 10; i++) {
            sb.append(para);
            try {
                //这里使用sleep，代替一个很慢的操作过程
                Thread.sleep(100);
            } catch (InterruptedException e) {
            }
        }
    }
}

```

```

        result =sb.toString();
    }
    public String getResult() {
        return result;
    }
}

```

接下来就是我们的客户端程序，Client主要实现了获取FutureData，并开启构造RealData的线程。并在接受请求后，很快的返回FutureData。注意，它不会等待数据真的构造完毕再返回，而是立即返回FutureData，即使这个时候FutureData内并没有真实数据。

```

public class Client {
    public Data request(final String queryStr) {
        final FutureData future = new FutureData();
        new Thread() {
            public void run() {
                // RealData的构建很耗时
                //所以在单独的线程中
                RealData realdata = new RealData(queryStr);
                future.setRealData(realdata);
            }
        }.start();
        return future;
        // FutureData会被立即返回
    }
}

```

最后，就是我们的主函数Main，它主要负责调用Client发起请求，并消费返回的数据。

```

public static void main(String[] args) {
    Client client = new Client();
    //这里会立即返回，因为得到的是FutureData而不是RealData
    Data data = client.request("name");
    System.out.println("请求完毕");
    try {
        //这里可以用一个sleep代替了对其他业务逻辑的处理
        //在处理这些业务逻辑的过程中，RealData被创建，从而充分利用了等待时间
        Thread.sleep(2000);
    } catch (InterruptedException e) {
    }
    //使用真实的数据
    System.out.println("数据 = " + data.getResult());
}

```

5.5.3 JDK中的Future模式

Future模式是如此常用，因此JDK内部已经为我们准备好了一套完整的实现。显然，这个实现要比我们前面提出的方案复杂得多。在这里，我们将简单向大家介绍一下它的使用方式。

首先，让我们看一下Future模式的基本结构，如图5.9所示。其中Future接口就类似于前文描述的订单或者说是契约。通过它，你可以得到真实的数据。RunnableFuture继承了Future和Runnable两个接口，其中run()方法用于构造真实的数据。它有一个具体的实现FutureTask类。FutureTask有一个内部类Sync，一些实质性的工作，会委托Sync类实

现。而Sync类最终会调用Callable接口，完成实际数据的组装工作。

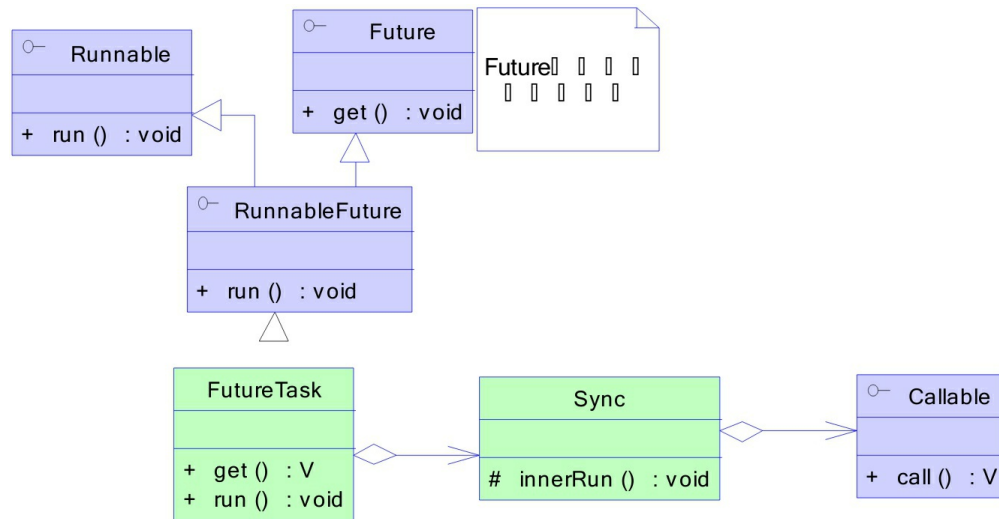


图5.9 JDK内置的Future模式

Callable接口只有一个方法call()，它会返回需要构造的实际数据。这个Callable接口也是这个Future框架和应用程序之间的重要接口。如果我们要实现自己的业务系统，通常需要实现自己的Callable对象。此外，FutureTask类也与应用密切相关，通常，我们会使用Callable实例构造一个FutureTask实例，并将它提交给线程池。

下面我们将展示这个内置的Future模式的使用：

```
01 public class RealData implements Callable<String> {
02     private String para;
03     public RealData(String para){
04         this.para=para;
05     }
06     @Override
07     public String call() throws Exception {
08
```

```

09     StringBuffer sb=new StringBuffer();
10     for (int i = 0; i < 10; i++) {
11         sb.append(para);
12         try {
13             Thread.sleep(100);
14         } catch (InterruptedException e) {
15         }
16     }
17     return sb.toString();
18 }
19 }

```

上述代码实现了Callable接口，它的call()方法会构造我们需要的真实数据并返回。当然这个过程可能是缓慢的，这里使用Thread.sleep()模拟它：

```

01 public class FutureMain {
02     public static void main(String[] args) throws InterruptedException
03         //构造FutureTask
04         FutureTask<String> future = new FutureTask<String>(r
05         ExecutorService executor = Executors.newFixedThreadPoo
06         //执行FutureTask，相当于上例中的 client.request("a") 发送请
07         //在这里开启线程进行RealData的call()执行
08         executor.submit(future);
09
10         System.out.println("请求完毕");
11         try {

```



```

12         //这里依然可以做额外的数据操作，这里使用sleep代替其他业务逻辑的
13         Thread.sleep(2000);
14     } catch (InterruptedException e) {
15     }
16     //相当于5.5.2节中得data.getResult ()，取得call()方法的返回值
17     //如果此时call()方法没有执行完成，则依然会等待
18     System.out.println("数据 = " + future.get());
19 }
20 }

```

上述代码就是使用Future模式的典型。第4行，构造了FutureTask对象实例，表示这个任务是有返回值的。构造FutureTask时，使用Callable接口，告诉FutureTask我们需要的数据应该如何产生。接着再第8行，将FutureTask提交给线程池。显然，作为一个简单的任务提交，这里必然是立即返回的，因此程序不会阻塞。接下来，我们不用关心数据是如何产生的。可以去做一些额外的事情，然后在需要的时候可以通过Future.get()（第18行）得到实际的数据。

除了基本的功能外，JDK还为Future接口提供了一些简单的控制功能：

```

boolean cancel(boolean mayInterruptIfRunning);           //取消任务
boolean isCancelled();                                     //是否已取消
boolean isDone();                                          //是否已完成
V get() throws InterruptedException, ExecutionException; //取得返回值
V get(long timeout, TimeUnit unit)                       //取得返回值

```

5.6 并行流水线

并发算法虽然可以充分发挥多核CPU的性能。但不幸的是，并非所有的计算都可以改造成并发的形式。那什么样的算法是无法使用并发进行计算的呢？简单来说，执行过程中有数据相关性的运算都是无法完美并行化的。

假如现在有两个数，B和C。如果我们要计算 $(B+C)*B/2$ ，那么这个运行过程就是无法并行的。原因是，如果B+C没有执行完成，则永远算不出 $(B+C)*B$ ，这就是数据相关性。如果线程执行时，所需的数据存在这种依赖关系，那么，就没有办法将它们完美的并行化。如图5.10所示，诠释了这个道理。

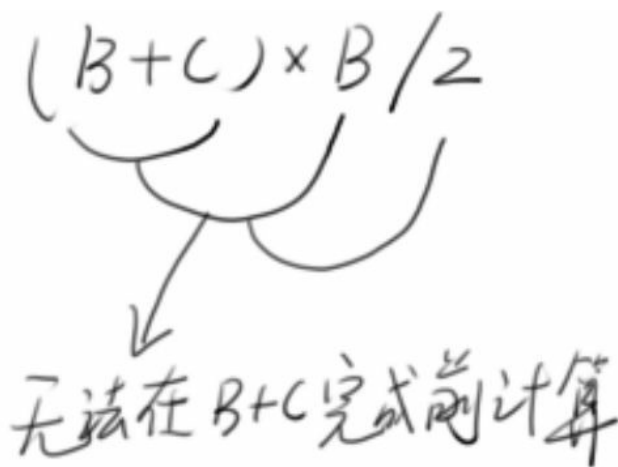


图5.10 $(B+C)*B/2$ 无法并行化

那遇到这种情况时，有没有什么补救措施呢？答案是肯定的，那就是借鉴日常生产中的流水线思想。

比如，现在要生产一批小玩偶。小玩偶的制作分为四个步骤，第一

要组装身体，第二要在身体上安装四肢和头部，第三，给组装完成的玩偶穿上一件漂亮的衣服，第四，就可以包装出货了。为了加快制作玩具的进度，我们不可能叫四个人同时加工一个玩具，因为这四个步骤有着严重的依赖关系。如果没有身体，就没有地方安装四肢，如果没有组装完成，就不能穿衣服，如果没有穿上衣服，就不能包装发货。因此，找四个人来做一个玩偶是毫无意义的。

但是，如果你现在要制作的不是1只玩偶，而是1万只玩偶，那情况就不同了。你可以找四个人，第一个人只负责组装身体，完成后交给第二个人；第二个人只负责安装头部和四肢，交付第三人；第三人只负责穿衣服，并交付第四人；第四人只负责包装发货。这样所有人都可以一起工作，共同完成任务，而整个时间周期也能缩短到原来的1/4左右，这就是流水线的思想。一旦流水线满载，每次只需要一步（假设一个玩偶需要四步）就可以产生一个玩偶，如图5.11所示。



图5.11 使用流水线生产玩偶

类似的思想可以借鉴到程序开发中。即使 $(B+C)*B/2$ 无法并行，但是如果你需要计算一大堆B和C的值，你依然可以将它流水化。首先将

计算过程拆分为三个步骤：

P1:A=B+C

P2:D=A×B

P3:D=D/2

上述步骤中P1、P2和P3均在单独的线程中计算，并且每个线程只负责自己的工作。此时，P3的计算结果就是最终需要的答案。

P1接收B和C的值，并求和，将结果输入给P2。P2求乘积后输入给P3。P3将D除以2得到最终值。一旦这条流水线建立，只需要一个计算步骤就可以得到(B+C)*B/2的结果。

为了实现这个功能，我们需要定义一个在线程间携带结果进行信息交换的载体：

```
public class Msg {  
    public double i;  
    public double j;  
    public String orgStr=null;  
}
```

P1计算的是加法：

```
01 public class Plus implements Runnable {  
02     public static BlockingQueue<Msg> bq=new LinkedBlockingQueue  
03     @Override  
04     public void run() {  
05         while(true){
```

```

06         try {
07             Msg msg=bq.take();
08             msg.j=msg.i+msg.j;
09             Multiply.bq.add(msg);
10         } catch (InterruptedException e) {
11         }
12     }
13 }
14 }

```

上述代码中，P1取得封装了两个操作数的Msg，并进行求和，将结果传递给乘法线程P2（第9行）。当没有数据需要处理时，P1进行等待。

P2计算乘法：

```

01 public class Multiply implements Runnable {
02     public static BlockingQueue<Msg> bq = new LinkedBlockingQueue<Msg>(10);
03
04     @Override
05     public void run() {
06         while (true) {
07             try {
08                 Msg msg = bq.take();
09                 msg.i = msg.i * msg.j;
10                 Div.bq.add(msg);
11             } catch (InterruptedException e) {
12             }
13         }
14     }
15 }

```

```
13     }
14 }
15 }
```

和P1非常类似，P2计算相乘结果后，将中间结果传递给除法线程P3。

P3计算除法：

```
01 public class Div implements Runnable {
02     public static BlockingQueue<Msg> bq = new LinkedBlockingQueue<Msg>(100000);
03
04     @Override
05     public void run() {
06         while (true) {
07             try {
08                 Msg msg = bq.take();
09                 msg.i = msg.i / 2;
10                 System.out.println(msg.orgStr + "=" + msg.i);
11             } catch (InterruptedException e) {
12             }
13         }
14     }
15 }
```

P3将结果除以2后输出最终的结果。

最后是提交任务的主线程，这里，我们提交100万个请求，让线程

组进行计算：

```
01 public class PStreamMain {
02     public static void main(String[] args) {
03         new Thread(new Plus()).start();
04         new Thread(new Multiply()).start();
05         new Thread(new Div()).start();
06
07         for (int i = 1; i <= 1000; i++) {
08             for (int j = 1; j <= 1000; j++) {
09                 Msg msg = new Msg();
10                 msg.i = i;
11                 msg.j = j;
12                 msg.orgStr = "(" + i + "+" + j + ")*" + i + "
13                 Plus.bq.add(msg);
14             }
15         }
16     }
17 }
```

上述代码第13行，将数据提交给P1加法线程，开启流水线的计算。在多核或者分布式场景中，这种设计思路可以有效地将有依赖关系的操作分配在不同的线程中进行计算，尽可能利用多核优势。

5.7 并行搜索

搜索是几乎每一个软件都必不可少的功能。对于有序数据，通常可以采用二分查找法。对于无序数据，则只能挨个查找。在本节中，我们将讨论有关并行的无序数组的搜索实现。

给定一个数组，我们要查找满足条件的元素。对于串行程序来说，只要遍历一下数组就可以得到结果。但如果要使用并行方式，则需要额外增加一些线程间的通信机制，使各个线程可以有效地运行。

一种简单的策略就是将原始数据集合按照期望的线程数进行分割。如果我们计划使用两个线程进行搜索，那么就可以把一个数组或集合分割成两个。每个线程各自独立搜索，当其中有一个线程找到数据后，立即返回结果即可。

现在假设有一个整型数组，我们需要查找数组内的元素：

```
static int[] arr;
```

定义线程池、线程数量以及存放结果的变量`result`。在`result`中，我们会保存符合条件的元素在`arr`数组中的下标。默认为-1，表示没有找到给定元素。

```
static ExecutorService pool = Executors.newCachedThreadPool();
static final int Thread_Num=2;
static AtomicInteger result=new AtomicInteger(-1);
```

并发搜索会要求每个线程查找`arr`中的一段，因此，搜索函数必须指

定线程需要搜索的起始和结束位置：

```
01 public static int search(int searchValue,int beginPos,int endP
02     int i=0;
03     for(i=beginPos;i<endPos;i++){
04         if(result.get()>=0){
05             return result.get();
06         }
07         if(arr[i] == searchValue){
08             //如果设置失败，表示其他线程已经先找到了
09             if(!result.compareAndSet(-1, i)){
10                 return result.get();
11             }
12             return i;
13         }
14     }
15     return -1;
16 }
```

上述代码第4行，首先通过`result`判断是否已经有其他线程找到了需要的结果。如果已经找到，则立即返回不再进行查找。如果没有找到，则进行下一步搜索。第7行代码成立则表示当前线程找到了需要的数据，那么就会将结果保存到`result`变量中。这里使用CAS操作，如果设置失败，则表示其他线程已经先我一步找到了结果。因此，可以无视失败的情况，找到结果后，进行返回。

定义一个线程进行查找，它会调用前面的`pSearch()`方法：

```

01 public static class SearchTask implements Callable<Integer>{
02     int begin,end,searchValue;
03     public SearchTask(int searchValue,int begin,int end){
04         this.begin=begin;
05         this.end=end;
06         this.searchValue=searchValue;
07     }
08     public Integer call(){
09         int re= search(searchValue,begin,end);
10         return re;
11     }
12 }

```

最后是pSearch()并行查找函数，它会根据线程数量对arr数组进行划分，并建立对应的任务提交给线程池处理：

```

01 public static int pSearch(int searchValue) throws InterruptedException
02 {
03     int subArrSize=arr.length/Thread_Num+1;
04     List<Future<Integer>> re=new ArrayList<Future<Integer>>();
05     for(int i=0;i<arr.length;i+=subArrSize){
06         int end = i+subArrSize;
07         if(end>=arr.length)end=arr.length;
08         re.add(pool.submit(new SearchTask(searchValue,i,end)));
09     }
10     for(Future<Integer> fu:re){
11         if(fu.get()>=0)return fu.get();
12     }
13 }

```

```
11     }  
12     return -1;  
13 }
```

上述代码中使用了JDK内置的Future模式，其中第4~8行将原始数组arr划分为若干段，并根据划分结果建立子任务。每一个子任务都会返回一个Future对象，通过Future对象可以获得线程组得到的最终结果。在这里，由于线程之间通过result共享彼此的信息，因此只要当一个线程成功返回后，其他线程都会立即返回。因此，不会出现由于排在前面的任务长时间无法结束而导致整个搜索结果无法立即获取的情况。

5.8 并行排序

排序是一项非常常用的操作。你的应用程序在运行时，可能无时无刻不在进行排序操作。排序的算法有很多，但在这里我并不打算一一介绍它们。对于大部分排序算法来说，都是串行执行的。当排序元素很多时，若使用并行算法代替串行算法，显然可以更加有效地利用CPU。但将串行算法改造成并行算法并非易事，甚至会极大地增加原有算法的复杂度。在这里，我将介绍几种相对简单的，但是也足以让人脑洞大开的平行排序算法。

5.8.1 分离数据相关性：奇偶交换排序

在介绍奇偶排序前，首先让我们看一下熟悉的冒泡排序。在这里，假设我们需要将数组进行从小到大的排序。冒泡排序的操作很类似水中的起泡上浮，在冒泡排序的执行过程中，如果数据较小，它就会逐步被交换到前面去，相反，对于大的数字，则会下沉，交换到数组的末尾。

冒泡排序的一般算法如下：

```
01 public static void bubbleSort(int[] arr) {  
02     for (int i = arr.length - 1; i > 0; i--) {  
03         for (int j = 0; j < i; j++) {  
04             if (arr[j] > arr[j + 1]) {  
05                 int temp = arr[j];
```

```

06         arr[j] = arr[j + 1];
07         arr[j + 1] = temp;
08     }
09 }
10 }
11 }

```

如图5.12所示，展示了冒泡排序的几次迭代过程：



图5.12 冒泡排序迭代过程

大家可以看到，在每次迭代的交换过程中，由于每次交换的两个元素存在数据冲突，对于每个元素，它既可能与前面的元素交换，也可能和后面的元素交换，因此很难直接改造成并行算法。

如果能够解开这种数据的相关性，就可以比较容易地使用并行算法

来实现类似的排序。奇偶交换排序就是基于这种思想的。

对于奇偶交换排序来说，它将排序过程分为两个阶段，奇交换和偶交换。对于奇交换来说，它总是比较奇数索引以及其相邻的后续元素。而偶交换总是比较偶数索引和其相邻的后续元素。并且，奇交换和偶交换会成对出现，这样才能保证比较和交换涉及到数组中的每一个元素。

奇偶交换的迭代示意图如图5.13所示。

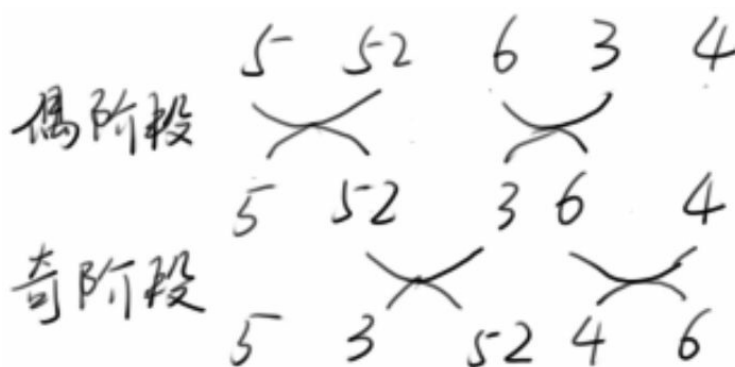


图5.13 奇偶交换迭代示意图

可以看到，由于将整个比较交换独立分割为奇阶段和偶阶段。这就使得在每一个阶段内，所有的比较和交换是没有数据相关性的。因此，每一次比较和交换都可以独立执行，也就可以并行化了。

下面是奇偶交换排序的串行实现：

```
01 public static void oddEvenSort(int[] arr) {  
02     int exchFlag = 1, start = 0;  
03     while (exchFlag == 1 || start == 1) {  
04         exchFlag = 0;  
05         for (int i = start; i < arr.length - 1; i += 2) {  
06             if (arr[i] > arr[i + 1]) {
```

```

07         int temp = arr[i];
08         arr[i] = arr[i + 1];
09         arr[i + 1] = temp;
10         exchFlag = 1;
11     }
12 }
13 if (start == 0)
14     start = 1;
15 else
16     start = 0;
17 }
18 }

```

其中，`exchFlag`用来记录当前迭代是否发生了数据交换，而`start`变量用来表示是奇交换还是偶交换。初始时，`start`为0，表示进行偶交换，每次迭代结束后，切换`start`的状态。如果上一次比较交换发生了数据交换，或者当前正在进行的是奇交换，循环就不会停止，直到程序不再发生交换，并且当前进行的是偶交换为止（表示奇偶交换已经成对出现）。

上述代码虽然是串行代码，但是已经可以很方便地改造成并行模式：

```

01 static int exchFlag=1;
02 static synchronized void setExchFlag(int v){
03     exchFlag=v;
04 }
05 static synchronized int getExchFlag(){

```

```
06     return exchFlag;
07 }
08
09 public static class OddEvenSortTask implements Runnable{
10     int i;
11     CountdownLatch latch;
12     public OddEvenSortTask(int i, CountdownLatch latch){
13         this.i=i;
14         this.latch=latch;
15     }
16     @Override
17     public void run() {
18         if (arr[i] > arr[i + 1]) {
19             int temp = arr[i];
20             arr[i] = arr[i + 1];
21             arr[i + 1] = temp;
22             setExchFlag(1);
23         }
24         latch.countDown();
25     }
26 }
27 public static void pOddEvenSort(int[] arr) throws InterruptedException
28     int start = 0;
29     while (getExchFlag() == 1 || start == 1) {
30         setExchFlag(0);
31         //偶数的数组长度，当start为1时，只有len/2-1个线程
32         CountdownLatch latch = new CountdownLatch(arr.length/2
```



```
33         for (int i = start; i < arr.length - 1; i += 2) {
34             pool.submit(new OddEvenSortTask(i,latch));
35         }
36         //等待所有线程结束
37         latch.await();
38         if (start == 0)
39             start = 1;
40         else
41             start = 0;
42     }
43 }
```

上述代码第9行，定义了奇偶排序的任务类。该任务的主要工作是进行数据比较和必要的交换（第18~23行）。并行排序的主体是 `pOddEvenSort()` 方法，它使用 `CountDownLatch` 记录线程数量，对于每一次迭代，使用单独的线程对每一次元素比较和交换进行操作。在下次迭代开始前，必须等待上一次迭代所有线程的完成。

5.8.2 改进的插入排序：希尔排序

插入排序也是一种很常用的排序算法。它的基本思想是：一个未排序的数组（当然也可以是链表）可以分为两个部分，前半部分是已经排序的，后半部分是未排序的。在进行排序时，只需要在未排序的部分中选择一个元素，将其插入到前面有序的数组中即可。最终，未排序的部分会越来越少，直到为0，那么排序就完成了。初始时，可以假设已排序部分就是第一个元素。

插入排序的几次迭代示意如图5.14所示。

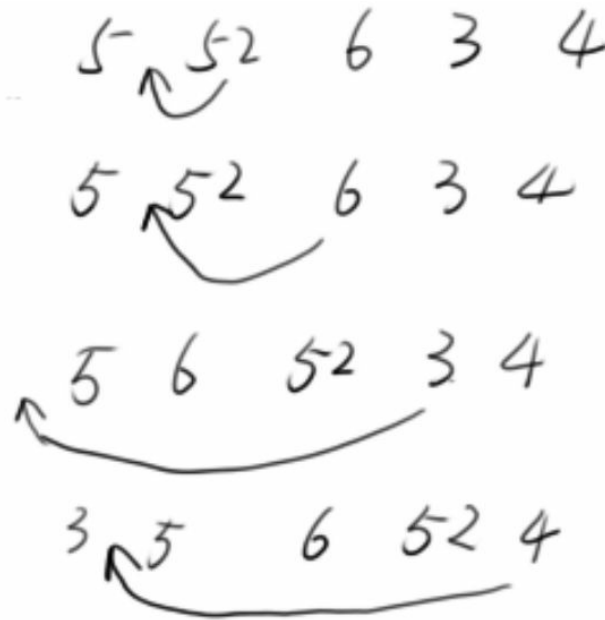


图5.14 插入排序示意图

插入排序的实现如下所示：

```
01 public static void insertSort(int[] arr) {  
02     int length = arr.length;  
03     int j, i, key;  
04     for (i = 1; i < length; i++) {  
05         //key为要准备插入的元素  
06         key = arr[i];  
07         j = i - 1;  
08         while (j >= 0 && arr[j] > key) {  
09             arr[j + 1] = arr[j];  
10             j--;  
11         }  
12         //找到合适的位置 插入key
```

```
13         arr[j + 1] = key;
14     }
15 }
```

上述代码第6行，提取要准备插入的元素（也就是未排序序列中的第一个元素）。接着，在已排序队列中找到这个元素的插入位置（第8～10行），并进行插入（第13行）即可。

简单的插入排序是很难并行化的。因为这一次的数据插入依赖于上一次得到的有序序列，因此多个步骤之间无法并行。为此，我们可以对插入排序进行扩展，这就是希尔排序。

希尔排序将整个数组根据间隔 h 分割为若干个子数组。子数组相互穿插在一起，每一次排序时，分别对每一个子数组进行排序。如图5.15所示，当 h 为3时，希尔排序将整个数组分为交织在一起的三个子数组。其中，所有的方块为一个子数组，所有的圆形、三角形分别组成另外两个子数组。每次排序时，总是交换间隔为 h 的两个元素。

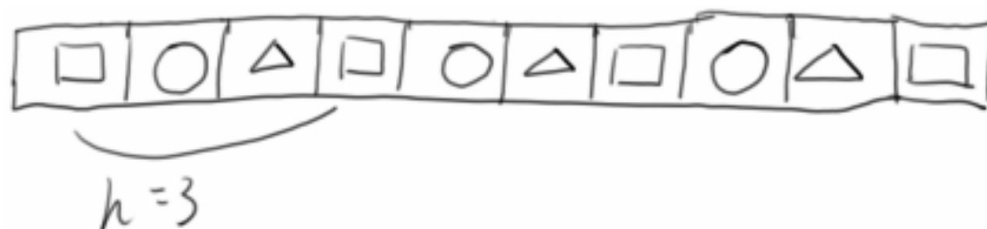


图5.15 $h=3$ 时的数组分割

在每一组排序完成后，可以递减 h 的值，进行下轮更加精细的排序。直到 h 为1，此时等价于一次插入排序。

希尔排序的一个主要优点是，即使一个较小的元素在数组的末尾，由于每次元素移动都以 h 为间隔进行，因此数组末尾的小元素可以在很

少的交换次数下，就被置换到最接近元素最终位置的地方。

下面是希尔排序的串行实现：

```
01 public static void shellSort(int[] arr) {
02     // 计算出最大的h值
03     int h = 1;
04     while (h <= arr.length / 3) {
05         h = h * 3 + 1;
06     }
07     while (h > 0) {
08         for (int i = h; i < arr.length; i++) {
09             if (arr[i] < arr[i - h]) {
10                 int tmp = arr[i];
11                 int j = i - h;
12                 while (j >= 0 && arr[j] > tmp) {
13                     arr[j + h] = arr[j];
14                     j -= h;
15                 }
16                 arr[j + h] = tmp;
17             }
18         }
19         // 计算出下一个h值
20         h = (h - 1) / 3;
21     }
22 }
```

上述代码第4~6行，计算一个合适的h值，接着正式进行希尔排

序。第8行的for循环进行间隔为h的插入排序，每次排序结束后，递减h的值（第20行）。直到h为1，退化为插入排序。

很显然，希尔排序每次都针对不同的子数组进行排序，各个子数组之间是完全独立的。因此，很容易改写成并程序序：

```
01 public static class ShellSortTask implements Runnable {
02     int i = 0;
03     int h = 0;
04     CountdownLatch l;
05
06     public ShellSortTask(int i, int h, CountdownLatch latch) {
07         this.i = i;
08         this.h = h;
09         this.l = latch;
10     }
11
12     @Override
13     public void run() {
14         if (arr[i] < arr[i - h]) {
15             int tmp = arr[i];
16             int j = i - h;
17             while (j >= 0 && arr[j] > tmp) {
18                 arr[j + h] = arr[j];
19                 j -= h;
20             }
21             arr[j + h] = tmp;
```

```
22     }
23     l.countDown();
24 }
25 }
26
27 public static void pShellSort(int[] arr) throws InterruptedException
28     // 计算出最大的h值
29     int h = 1;
30     CountdownLatch latch = null;
31     while (h <= arr.length / 3) {
32         h = h * 3 + 1;
33     }
34     while (h > 0) {
35         System.out.println("h=" + h);
36         if (h >= 4)
37             latch = new CountdownLatch(arr.length - h);
38         for (int i = h; i < arr.length; i++) {
39             // 控制线程数量
40             if (h >= 4) {
41                 pool.execute(new ShellSortTask(i, h, latch));
42             } else {
43                 if (arr[i] < arr[i - h]) {
44                     int tmp = arr[i];
45                     int j = i - h;
46                     while (j >= 0 && arr[j] > tmp) {
47                         arr[j + h] = arr[j];
48                         j -= h;
```

```
49         }
50         arr[j + h] = tmp;
51     }
52     // System.out.println(Arrays.toString(arr));
53 }
54 }
55 // 等待线程排序完成，进入下一次排序
56 latch.await();
57 // 计算出下一个h值
58 h = (h - 1) / 3;
59 }
60 }
```

上述代码中定义ShellSortTask作为并行任务。一个ShellSortTask的作用是根据给定的起始位置和h，对子数组进行排序，因此可以完全并行化。

为控制线程数量，这里定义并行主函数pShellSort()在h大于或等于4时使用并行线程（第40行），否则则退化为传统的插入排序。

每次计算后，递减h的值（第58行）。

5.9 并行算法：矩阵乘法

我在第一章中已经提到，Linus认为并行程序目前只有在服务端程序和图像处理领域有发展的空间。且不论这种说法是否正确，但从中也可以看出并发对于这两个应用领域的重要性。而对于图像处理来说，矩阵运行是其中必不可少的重要数学方法。当然，除了图像处理，矩阵运算在神经网络、模式识别等领域也有着广泛的用途。在这里，我将向大家介绍矩阵运算的典型代表——矩阵乘法的并行化实现。

在矩阵乘法中，第一个矩阵的列数和第二个矩阵的行数必须是相同的。如图5.16所示，矩阵A和矩阵B相乘，其中矩阵A为4行2列，矩阵B为2行4列，它们相乘后，得到的是4行4列的矩阵，并且新矩阵中每一个元素为矩阵A和B对应行列的乘积求和。

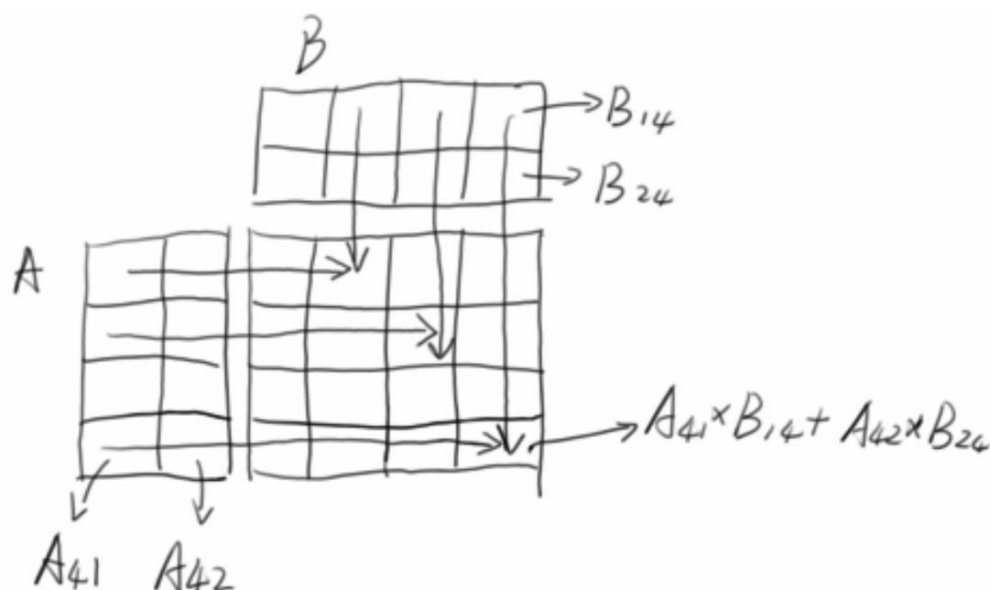


图5.16 矩阵相乘示意图

如果需要进行并行计算，一种简单的策略是可以将A矩阵进行水平

分割，得到子矩阵 A_1 和 A_2 ， B 矩阵进行垂直分割，得到子矩阵 B_1 和 B_2 。此时，我们只要分别计算这些子矩阵的乘积，将结果进行拼接，就能得到原始矩阵 A 和 B 的乘积。如图5.17所示，展示了这种并行计算的策略。

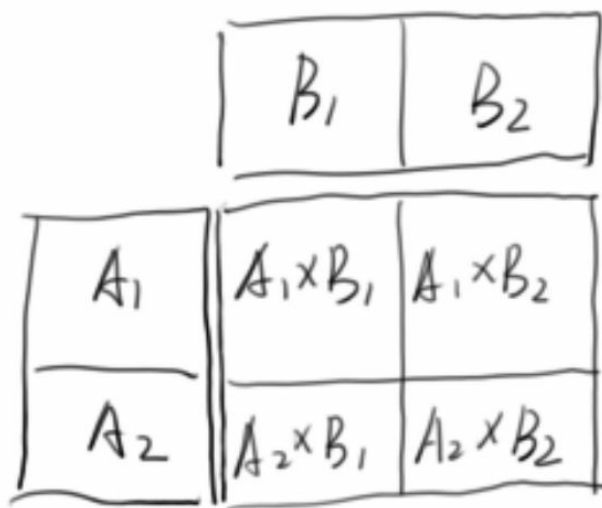


图5.17 矩阵拆分进行并行计算

当然，这个过程是可以反复进行的。为了计算 $A_1 \times B_1$ ，我们还可以进一步将 A_1 和 B_1 进行分解，直到我们认为子矩阵的大小已经在可接受范围内。

这里，我们使用ForkJoin框架来实现这个并行矩阵相乘的想法。为了方便矩阵计算，我们使用jMatrices开源软件，作为矩阵计算的工具。其中，使用的主要API如下：

- **Matrix**: 代表一个矩阵
- **MatrixOperator.multiply(Matrix, Matrix)**: 矩阵相乘
- **Matrix.row()**: 获得矩阵的行数
- **Matrix.getSubMatrix()**: 获得矩阵的子矩阵
- **MatrixOperator.horizontalConcatenation(Matrix, Matrix)**: 将两个矩

阵进行水平连接

- `MatrixOperator.verticalConcatenation(Matrix,Matrix)`: 将两个矩阵进行垂直连接

为了计算矩阵乘法，定义一个任务类`MatrixMulTask`。它会进行矩阵相乘的计算，如果输入矩阵的粒度比较大，则会再次进行任务分解：

```
01 public class MatrixMulTask extends RecursiveTask<Matrix> {
02     Matrix m1;
03     Matrix m2;
04     String pos;
05
06     public MatrixMulTask(Matrix m1, Matrix m2, String pos) {
07         this.m1 = m1;
08         this.m2 = m2;
09         this.pos = pos;
10     }
11
12     @Override
13     protected Matrix compute() {
14         //System.out.println(Thread.currentThread().getId()+" :
15         getName() + " is start");
16         if (m1.rows() <= PMatrixMul.granularity || m2.cols() <= PMatrixMul.granularity) {
17             Matrix mRe = MatrixOperator.multiply(m1, m2);
18             return mRe;
19         } else {
20             // 如果不是，那么继续分割矩阵
```

```
20         int rows;
21         rows = m1.rows();
22         // 左乘的矩阵横向分割
23         Matrix m11 = m1.getSubMatrix(1, 1, rows / 2, m1.co
24         Matrix m12 = m1.getSubMatrix(rows / 2 + 1, 1, m1.r
25         // 右乘矩阵纵向分割
26         Matrix m21 = m2.getSubMatrix(1, 1, m2.rows(), m2.c
27         Matrix m22 = m2.getSubMatrix(1, m2.cols() / 2 + 1,
28
29         ArrayList<MatrixMulTask> subTasks = new ArrayList
30         MatrixMulTask tmp = null;
31         tmp = new MatrixMulTask(m11, m21, "m1");
32         subTasks.add(tmp);
33         tmp = new MatrixMulTask(m11, m22, "m2");
34         subTasks.add(tmp);
35         tmp = new MatrixMulTask(m12, m21, "m3");
36         subTasks.add(tmp);
37         tmp = new MatrixMulTask(m12, m22, "m4");
38         subTasks.add(tmp);
39         for (MatrixMulTask t : subTasks) {
40             t.fork();
41         }
42         Map<String, Matrix> matrixMap = new HashMap<Stri
43         for (MatrixMulTask t : subTasks) {
44             matrixMap.put(t.pos, t.join());
45         }
46         Matrix tmp1 = MatrixOperator.horizontalConcatenati
```

```

matrixMap.get("m2"));
47         Matrix tmp2 = MatrixOperator.horizontalConcatenati
matrixMap.get("m4"));
48         Matrix reM = MatrixOperator.verticalConcatenation(
49         return reM;
50     }
51 }
52 }

```

`MatrixMulTask`类由三个参数构成，分别是需要计算的矩阵双方，以及计算结果位于父矩阵相乘结果中的位置，如图5.18所示。

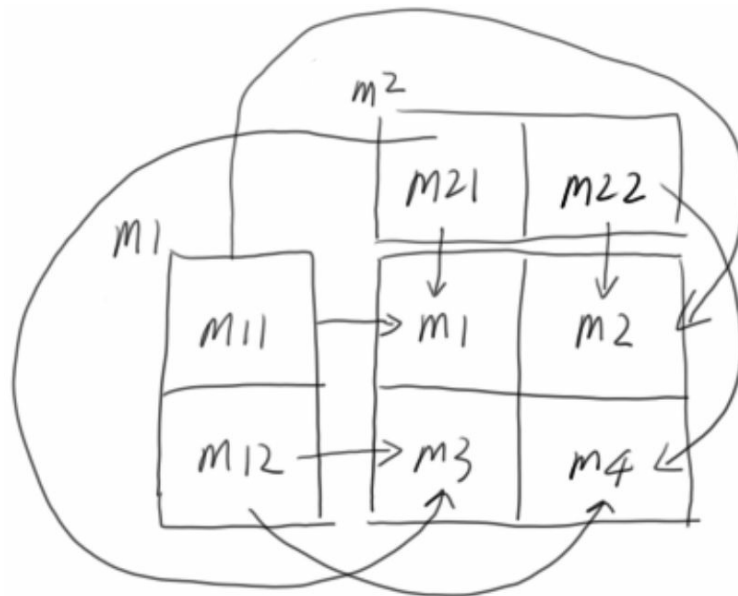


图5.18 矩阵分解方式

`MatrixMulTask`中的成员变量`m1`和`m2`表示要相乘的两个矩阵，`pos`表示这个乘积结果在父矩阵相乘结果中所处的位置，有`m1`、`m2`、`m3`和`m4`等四种。代码第23~27行先对矩阵进行分割，分割后得到`m11`、`m12`、`m21`和`m22`等四个矩阵，并将它们按照如图5.18所示的规则进行子

任务的创建。在第39~41行，计算这些子任务。在子任务返回后，在第42~48行将返回的四个矩阵m1、m2、m3和m4拼接成新的矩阵作为最终结果。

如果矩阵的粒度足够小就直接进行运算而不进行分解（第16行）。

使用这个任务类可以很容易地进行矩阵并行运算，下面是使用方法：

```
01 public static final int granularity=3;
02 public static void main(String[] args) throws InterruptedException
03 {
04     ForkJoinPool forkJoinPool = new ForkJoinPool();
05     Matrix m1=MatrixFactory.getRandomIntMatrix(300, 300, null)
06     Matrix m2=MatrixFactory.getRandomIntMatrix(300, 300, null)
07     MatrixMulTask task=new MatrixMulTask(m1,m2,null);
08     ForkJoinTask<Matrix> result = forkJoinPool.submit(task);
09     Matrix pr=result.get();
10     System.out.println(pr);
11 }
```

上述代码中第4~5行创建两个300*300的随机矩阵。构造矩阵计算任务MatrixMulTask并将其提交给ForkJoinPool线程池。第8行执行ForkJoinTask.get()方法等待并获得最终结果。

5.10 准备好了再通知我：网络NIO

Java NIO是New IO的简称，它是一种可以替代Java IO的一套新的IO机制。它提供了一套不同于Java标准IO的操作机制。严格来说，NIO与并发并无直接的关系。但是，使用NIO技术可以大大提高线程的使用效率。

Java NIO中涉及的基础内容有通道（Channel）和缓冲区（Buffer）、文件IO和网络IO。有关通道、缓冲区以及文件IO在这里不打算进行详细的介绍，大家可以参考本章的参考文献。在这里，我想多花一点时间详细介绍一下有关网络IO的内容。

对于标准的网络IO来说，我们会使用Socket进行网络的读写。为了让服务器可以支持更多的客户端连接，通常的做法是为每一个客户端连接开启一个线程。让我们先回顾一下这方面的内容。

5.10.1 基于Socket的服务端的多线程模式

这里，我以一个简单的Echo服务器为例。对于Echo服务器，它会读取客户端的一个输入，并将这个输入原封不动地返回给客户端。这看起来很简单，但是麻雀虽小五脏俱全。为了完成这个功能，服务器还是需要有一套完整的Socket处理机制。因此，这个Echo服务器非常适合来进

行学习。实际上，我认为任何业务逻辑简单的系统都很适合学习，大家不用为了去理解业务上复杂的功能而忽略了系统的重点。

服务端使用多线程进行处理时的结构示意图，如图5.19所示。

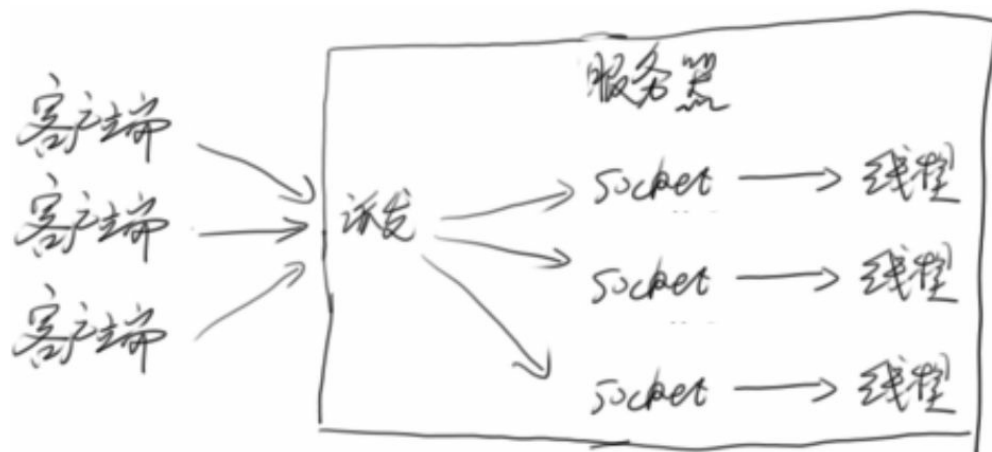


图5.19 多线程的服务端

服务器会为每一个客户端连接启用一个线程，这个新的线程将全心全意为这个客户端服务。同时，为了接受客户端连接，服务器还会额外使用一个派发线程。

下面的代码实现了这个服务器：

```
01 public class MultiThreadEchoServer {
02     private static ExecutorService tp=Executors.newCachedThre
03     static class HandleMsg implements Runnable{
04         Socket clientSocket;
05         public HandleMsg(Socket clientSocket){
06             this.clientSocket=clientSocket;
07         }
08 }
```

```
09     public void run(){
10         BufferedReader is =null;
11         PrintWriter os = null;
12         try {
13
14             is = new BufferedReader(new InputStreamReader(clientSo
15                 os = new PrintWriter(clientSocket.getOutputStream()
16                 // 从InputStream当中读取客户端所发送的数据
17                 String inputLine = null;
18                 long b=System.currentTimeMillis();
19                 while ((inputLine = is.readLine()) != null) {
20                     os.println(inputLine);
21                 }
22                 long e=System.currentTimeMillis();
23                 System.out.println("spend: "+(e-b)+"ms");
24             } catch (IOException e) {
25                 e.printStackTrace();
26             }finally{
27                 try {
28                     if(is!=null)is.close();
29                     if(os!=null)os.close();
30                     clientSocket.close();
31                 } catch (IOException e) {
32                     e.printStackTrace();
33                 }
34             }
35         }
```



```

36     }
37     public static void main(String args[]) {
38         ServerSocket echoServer = null;
39         Socket clientSocket = null;
40         try {
41             echoServer = new ServerSocket(8000);
42         } catch (IOException e) {
43             System.out.println(e);
44         }
45         while (true) {
46             try {
47                 clientSocket = echoServer.accept();
48                 System.out.println(clientSocket.getRemoteSocket().getInetAddress().getHostAddress());
49                 tp.execute(new HandleMsg(clientSocket));
50             } catch (IOException e) {
51                 System.out.println(e);
52             }
53         }
54     }
55 }

```

第2行，我们使用了一个线程池来处理每一个客户端连接。第3~33行，定义了HandleMsg线程，它由一个客户端Socket构造而成，它的任务是读取这个Socket的内容并将其进行返回，返回成功后，任务完成，客户端Socket就被正常关闭。其中第23行，统计并输出了服务端线程处理一次客户端请求所花费的时间（包括读取数据和回写数据的时间）。主线程main的主要作用是在8000端口上进行等待。一旦有新的客户端连

接，它会根据这个连接创建HandleMsg线程进行处理（第47~49行）。

这就是一个支持多线程的服务端的核心内容。它的特点是，在相同可支持的线程范围内，可以尽量多地支持客户端的数量，同时和单线程服务器相比，它也可以更好地使用多核CPU。

为了方便大家学习，这里再给出一个客户端的参考实现：

```
01 public static void main(String[] args) throws IOException {
02     Socket client = null;
03     PrintWriter writer = null;
04     BufferedReader reader = null;
05     try {
06         client = new Socket();
07         client.connect(new InetSocketAddress("localhost", 8000));
08         writer = new PrintWriter(client.getOutputStream(), true);
09         writer.println("Hello!");
10         writer.flush();
11
12         reader = new BufferedReader(new InputStreamReader(client.getInputStream()));
13         System.out.println("from server: " + reader.readLine());
14     } catch (UnknownHostException e) {
15         e.printStackTrace();
16     } catch (IOException e) {
17         e.printStackTrace();
18     } finally {
19         if (writer != null)
20             writer.close();
    }
```

```
21         if (reader != null)
22             reader.close();
23         if (client != null)
24             client.close();
25     }
26 }
```

上述代码在第7行，连接了服务器的8000端口，并发送字符串。接着在第12行，读取服务器的返回信息并进行输出。

可以说，这种多线程的服务器开发模式是极其常用的。对于绝大多数应用来说，这种模式可以很好地工作。但是，如果你想让你的程序工作得更加有效，就必须知道这种模式的一个重大弱点——那就是它倾向于让CPU进行IO等待。为了理解这一点，让我们看一下下面这个比较极端的例子：

```
01 public class HeavySocketClient {
02     private static ExecutorService tp=Executors.newCachedThre
03     private static final int sleep_time=1000*1000*1000;
04     public static class EchoClient implements Runnable{
05         public void run(){
06             Socket client = null;
07             PrintWriter writer = null;
08             BufferedReader reader = null;
09             try {
10                 client = new Socket();
11                 client.connect(new InetSocketAddress("localhost
12                 writer = new PrintWriter(client.getOutputStream
```

```
13         writer.print("H");
14         LockSupport.parkNanos(sleep_time);
15         writer.print("e");
16         LockSupport.parkNanos(sleep_time);
17         writer.print("l");
18         LockSupport.parkNanos(sleep_time);
19         writer.print("l");
20         LockSupport.parkNanos(sleep_time);
21         writer.print("o");
22         LockSupport.parkNanos(sleep_time);
23         writer.print("!");
24         LockSupport.parkNanos(sleep_time);
25         writer.println();
26         writer.flush();
27
28         reader = new BufferedReader(new InputStreamReader
29             System.out.println("from server: " + reader.re
30     } catch (UnknownHostException e) {
31         e.printStackTrace();
32     } catch (IOException e) {
33         e.printStackTrace();
34     } finally {
35         try {
36             if (writer != null)
37                 writer.close();
38             if (reader != null)
39                 reader.close();
```

```

40             if (client != null)
41                 client.close();
42         } catch (IOException e) {
43         }
44     }
45 }
46 }
47 public static void main(String[] args) throws IOException
48     EchoClient ec=new EchoClient();
49     for(int i=0;i<10;i++)
50         tp.execute(ec);
51 }
52 }

```

上述代码定义了一个新的客户端，它会进行10次请求（第49～50行开启10个线程）。每一次请求都会访问8000端口。连接成功后，会向服务器输出“Hello!”字符串（第13～26行），但是在这一次交互中，客户端会慢慢地进行输出，每次只输出一个字符，之后进行1秒的等待。因此，整个过程会持续6秒。

开启多线程池的服务器和上述客户端。服务器端的部分输出如下：

```

spend:6000ms
spend:6000ms
spend:6000ms
spend:6001ms
spend:6002ms
spend:6002ms

```

```
spend:6002ms  
spend:6002ms  
spend:6003ms  
spend:6003ms
```

可以看到，对于服务端来说，每一个请求的处理时间都在6秒左右。这很容易理解，因为服务器要先读入客户端的输入，而客户端缓慢的处理速度（当然也可能是一个拥塞的网络环境）使得服务器花费了不少等待时间。

我们可以试想一下，如果服务器要处理大量的请求连接，每个请求如果都像这样拖慢了服务器的处理速度，那么服务端能够处理的并发数量就会大幅度减少。反之，如果服务器每次都能很快地处理一次请求，那么相对的，它的并发能力就能上升。

在这个案例中，服务器处理请求之所以慢，并不是因为在服务端有多少繁重的任务，而仅仅是因为服务线程在等待IO而已。让高速运转的CPU去等待极其低效的网络IO是非常不合算的行为。那么，我们是不是可以想一个方法，将网络IO的等待时间从线程中分离出来呢？

5.10.2 使用NIO进行网络编程

使用Java的NIO就可以将上面的网络IO等待时间从业务处理线程中抽取出来。那么NIO是什么，它又是如何工作的呢？

要了解NIO，我们首先需要知道在NIO中的一个关键组件Channel（通道）。Channel有点类似于流，一个Channel可以和文件或者

网络Socket对应。如果Channel对应着一个Socket，那么往这个Channel中写数据，就等同于向Socket中写入数据。

和Channel一起使用的另外一个重要组件就是Buffer。大家可以简单地把Buffer理解成一个内存区域或者byte数组。数据需要包装成Buffer的形式才能和Channel交互（写入或者读取）。

另外一个与Channel密切相关的是Selector（选择器）。在Channel的众多实现中，有一个SelectableChannel实现，表示可被选择的通道。任何一个SelectableChannel都可以将自己注册到一个Selector中。这样，这个Channel就能被Selector所管理。而一个Selector可以管理多个SelectableChannel。当SelectableChannel的数据准备好时，Selector就会接到通知，得到那些已经准备好的数据。而SocketChannel就是SelectableChannel的一种。因此，它们构成了如图5.20所示的结构。

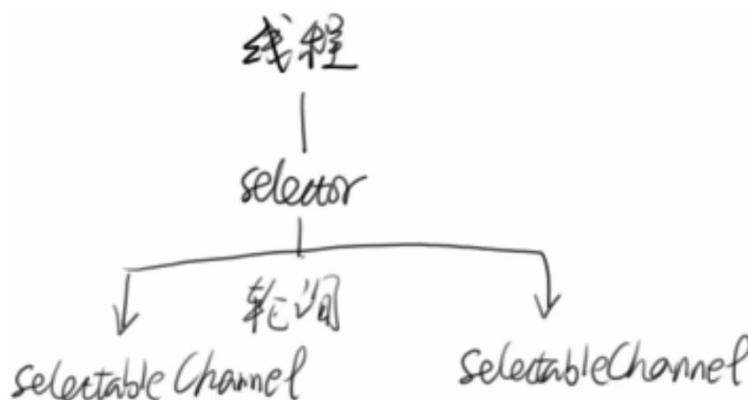


图5.20 Selector和Channel

大家可以看到，一个Selector可以由一个线程进行管理，而一个SocketChannel则可以表示一个客户端连接，因此这就构成由一个或者极少数线程，来处理大量客户端连接的结构。当与客户端连接的数据没有准备好时，Selector会处于等待状态（不过幸好，用于管理Selector的线程数是极少量的），而一旦有任何一个SocketChannel准备好了数据，

Selector就能立即得到通知，获取数据进行处理。

下面就让我们用NIO来重新构造这个多线程的Echo服务器吧！

首先，我们需要定义一个Selector和线程池：

```
private Selector selector;  
private ExecutorService tp=Executors.newCachedThreadPool();
```

其中，selector用于处理所有的网络连接。线程池tp用于对每一个客户端进行相应的处理，每一个请求都会委托给线程池中的线程进行实际的处理。

为了能够统计服务器线程在一个客户端上花费了多少时间，这里还需要定义一个与时间统计有关的类：

```
public static Map<Socket,Long> time_stat=new HashMap<Socket,Long>();
```

它用于统计在某一个Socket上花费的时间，time_stat的key为Socket，value为时间戳（可以记录处理开始时间）。

下面就可以来看一下NIO服务器的核心代码，下面的startServer()方法用于启动NIO Server：

```
01 private void startServer() throws Exception {  
02     selector = SelectorProvider.provider().openSelector();  
03     ServerSocketChannel ssc = ServerSocketChannel.open();  
04     ssc.configureBlocking(false);  
05  
06     InetSocketAddress isa = new InetSocketAddress(InetAddress.
```



```
07     InetSocketAddress isa = new InetSocketAddress(8000);
08     ssc.socket().bind(isa);
09
10     SelectionKey acceptKey = ssc.register(selector, SelectionK
11
12     for (;;) {
13         selector.select();
14         Set readyKeys = selector.selectedKeys();
15         Iterator i = readyKeys.iterator();
16         long e=0;
17         while (i.hasNext()) {
18             SelectionKey sk = (SelectionKey) i.next();
19             i.remove();
20
21             if (sk.isAcceptable()) {
22                 doAccept(sk);
23             }
24             else if (sk.isValid() && sk.isReadable()) {
25                 if(!time_stat.containsKey(((SocketChannel)sk.c
26                     time_stat.put(((SocketChannel)sk.channel()
27                         System.currentTimeMillis()));
28                 doRead(sk);
29             }
30             else if (sk.isValid() && sk.isWritable()) {
31                 doWrite(sk);
32                 e=System.currentTimeMillis();
33                 long b=time_stat.remove(((SocketChannel)sk.cha
```

```
34             System.out.println("spend: "+(e-b)+"ms");
35         }
36     }
37 }
38 }
```

上述代码第2行，通过工厂方法获得一个Selector对象的实例。第3行，获得表示服务端的SocketChannel实例。第4行，将这个SocketChannel设置为非阻塞模式。实际上，Channel也可以像传统的Socket那样按照阻塞的方式工作。但在这里，更倾向于让其工作在非阻塞模式，在这种模式下，我们才可以向Channel注册感兴趣的事件，并且在数据准备好时，得到必要的通知。接着，在第6~8行进行端口绑定，将这个Channel绑定在8000端口。

在第10行，将这个ServerSocketChannel绑定到Selector上，并注册它感兴趣的时间为Accept。这样，Selector就能为这个Channel服务了。当Selector发现ServerSocketChannel有新的客户端连接时，就会通知ServerSocketChannel进行处理。方法register()的返回值是一个SelectionKey，SelectionKey表示一对Selector和Channel的关系。当Channel注册到Selector上时，就相当于确立了两者的服务关系，那么SelectionKey就是这个契约。当Selector或者Channel被关闭时，它们对应的SelectionKey就会失效。

第12~37行是一个无穷循环，它的主要任务就是等待-分发网络消息。

第13行的select()方法是一个阻塞方法。如果当前没有任何数据准备好，它就会等待。一旦有数据可读，它就会返回。它的返回值是已经准

备就绪的SelectionKey的数量。这里简单地将其忽略。

第14行获取那些准备好的SelectionKey。因为Selector同时为多个Channel服务，因此已经准备就绪的Channel就有可能是多个。所以，这里得到的自然是一个集合。得到这个就绪集合后，剩下的就是遍历这个集合，挨个处理所有的Channel数据。

第15行得到这个集合的迭代器。第17行使用迭代器遍历整个集合。第18行根据迭代器获得一个集合内的SelectionKey实例。

第19行将这个元素移除！注意，这个非常重要，否则就会重复处理相同的SelectionKey。当你处理完一个SelectionKey后，务必将其从集合内删除。

第21行判断当前SelectionKey所代表的Channel是否在Acceptable状态，如果是，就进行客户端的接收（执行doAccept()方法）。

第24行判断Channel是否已经可以读了，如果是就进行读取（doRead()方法）。这里为了统计系统处理每一个连接的时间，在第25～27行记录了在读取数据之前的一个时间戳。

第30行判断通道是否准备好进行写。如果是就进行写入（doWrite()方法），同时在写入完成后，根据读取前的时间戳，输出处理这个Socket连接的耗时。

在了解服务端的整体框架后，下面让我们从细节着手，学习一下几个主要方法的内部实现。首先是doAccept()方法，它与客户端建立连接：

```
01 private void doAccept(SelectionKey sk) {
```

```
02     ServerSocketChannel server = (ServerSocketChannel) sk.chan
03     SocketChannel clientChannel;
04     try {
05         clientChannel = server.accept();
06         clientChannel.configureBlocking(false);
07
08         // Register this channel for reading.
09         SelectionKey clientKey = clientChannel.register(select
10         // Allocate an EchoClient instance and attach it to th
11         EchoClient echoClient = new EchoClient();
12         clientKey.attach(echoClient);
13
14         InetAddress clientAddress = clientChannel.socket().get
15         System.out.println("Accepted connection from " + clien
16     } catch (Exception e) {
17         System.out.println("Failed to accept new client.");
18         e.printStackTrace();
19     }
20 }
```

和Socket编程很类似，当有一个新的客户端连接接入时，就会有一个新的Channel产生代表这个连接。上述代码第5行，生成的clientChannel就表示和客户端通信的通道。第6行，将这个Channel配置为非阻塞模式，也就是要求系统在准备好IO后，再通知我们的线程来读取或者写入。

第9行很关键，它将新生成的Channel注册到selector选择器上，并告

诉Selector，我现在对读（OP_READ）操作感兴趣。这样，当Selector发现这个Channel已经准备好读时，就能给线程一个通知。

第11行新建一个对象实例，一个EchoClient实例代表一个客户端。在第12行，我们将这个客户端实例作为附件，附加到表示这个连接的SelectionKey上。这样在整个连接的处理过程中，我们都可以共享这个EchoClient实例。

EchoClient的定义很简单，它封装了一个队列，保存在需要回复给这个客户端的所有信息，这样，再进行回复时，只要从outq对象中弹出元素即可。

```
class EchoClient {
    private LinkedList<ByteBuffer> outq;
    EchoClient() {
        outq = new LinkedList<ByteBuffer>();
    }
    public LinkedList<ByteBuffer> getOutputQueue() {
        return outq;
    }
    public void enqueue(ByteBuffer bb) {
        outq.addFirst(bb);
    }
}
```

下面来看一下另外一个重要的方法doRead()。当Channel可以读取时，doRead()方法就会被调用。

```
01 private void doRead(SelectionKey sk) {
02     SocketChannel channel = (SocketChannel) sk.channel();
03     ByteBuffer bb = ByteBuffer.allocate(8192);
04     int len;
05
06     try {
07         len = channel.read(bb);
08         if (len < 0) {
09             disconnect(sk);
10             return;
11         }
12     } catch (Exception e) {
13         System.out.println("Failed to read from client.");
14         e.printStackTrace();
15         disconnect(sk);
16         return;
17     }
18
19     bb.flip();
20     tp.execute(new HandleMsg(sk, bb));
21 }
```

方法doRead()接收一个SelectionKey参数，通过这个SelectionKey可以得到当前的客户端Channel（第2行）。在这里，我们准备8K的缓冲区读取数据，所有读取的数据存放在变量bb中（第7行）。读取完成后，重置缓冲区，为数据处理做准备（第19行）。

在这个示例中，我们对数据的处理很简单。但是为了模拟复杂的场景，还是使用了线程池进行数据处理（第20行）。这样，如果数据处理很复杂，就能在单独的线程中进行，而不用阻塞任务派发线程。

HandleMsg的实现也很简单：

```
01 class HandleMsg implements Runnable{
02     SelectionKey sk;
03     ByteBuffer bb;
04     public HandleMsg(SelectionKey sk,ByteBuffer bb){
05         this.sk=sk;
06         this.bb=bb;
07     }
08     @Override
09     public void run() {
10         EchoClient echoClient = (EchoClient) sk.attachment();
11         echoClient.enqueue(bb);
12         sk.interestOps(SelectionKey.OP_READ | SelectionKey.OP_
13         //强迫selector立即返回
14         selector.wakeup();
15     }
16 }
```

上述代码，简单地将接收到的数据压入EchoClient的队列（第11行）。如果需要处理业务逻辑，就可以在这里进行处理。

在数据处理完成后，就可以准备将结果回写到客户端，因此，重新注册感兴趣的消息事件，将写操作（OP_WRITE）也作为感兴趣的事件

进行提交（第12行）。这样在通道准备好写入时，就能通知线程。

写入操作使用doWrite()函数实现：

```
01 private void doWrite(SelectionKey sk) {
02     SocketChannel channel = (SocketChannel) sk.channel();
03     EchoClient echoClient = (EchoClient) sk.attachment();
04     LinkedList<ByteBuffer> outq = echoClient.getOutputQueue()
05
06     ByteBuffer bb = outq.getLast();
07     try {
08         int len = channel.write(bb);
09         if (len == -1) {
10             disconnect(sk);
11             return;
12         }
13
14         if (bb.remaining() == 0) {
15             // The buffer was completely written, remove it.
16             outq.removeLast();
17         }
18     } catch (Exception e) {
19         System.out.println("Failed to write to client.");
20         e.printStackTrace();
21         disconnect(sk);
22     }
23 }
```



```
24     if (outq.size() == 0) {
25         sk.interestOps(SelectionKey.OP_READ);
26     }
27 }
```

函数doWrite()也接收一个SelectionKey，当然针对一个客户端来说，这个SelectionKey实例和doRead()拿到的SelectionKey是同一个。因此，通过SelectionKey我们就可以在这两个操作中共享EchoClient实例。上述代码第3~4行，我们取得EchoClient实例以及它的发送内容列表。第6行，获得列表顶部元素，准备写回客户端。第8行进行写回操作。如果全部发送完成，则移除这个缓存对象（第16行）。

在doWrite()中最重要的，也是最容易被忽略的是在全部数据发送完成后（也就是outq的长度为0），需要将写事件（OP_WRITE）从感兴趣的操作中移除（第25行）。如果不这么做，每次Channel准备好写时，都会来执行doWrite()方法。而实际上，你又无数据可写，这显然是不合理的。因此，这个操作很重要。

上面我们已经介绍了主要的核心代码，现在使用这个NIO服务器来处理上一节中客户端的访问。同样的，客户端也是要花费将近6秒钟，才能完成一次消息的发送，那使用NIO技术后，服务端线程需要花费多少时间来处理这些请求呢？答案如下：

```
spend:2ms
spend:2ms
spend:2ms
spend:2ms
spend:3ms
```

```
spend:3ms  
spend:0ms  
spend:0ms  
spend:2ms  
spend:3ms
```

可以看到，在使用NIO技术后，即使客户端迟钝或者出现了网络延迟等现象，并不会给服务器带来太大的问题。

5.10.3 使用NIO来实现客户端

在前面的案例中，我们使用Socket编程来构建我们的客户端，使用NIO来实现服务端。实际上，使用NIO也可以用来创建客户端。这里，我们再演示一下使用NIO创建客户端的例子。

和构造服务器类似，核心的元素也是Selector、Channel和SelectionKey。

首先，我们需要初始化Selector和Channel：

```
01 private Selector selector;  
02 public void init(String ip, int port) throws IOException {  
03     SocketChannel channel = SocketChannel.open();  
04     channel.configureBlocking(false);  
05     this.selector = SelectorProvider.provider().openSelector();  
06     channel.connect(new InetSocketAddress(ip, port));  
07     channel.register(selector, SelectionKey.OP_CONNECT);
```

08 }

上述代码第3行，创建一个SocketChannel实例，并设置为非阻塞模式。第5行创建了一个Selector。第6行，将SocketChannel绑定到Socket上。但由于当前Channel是非阻塞的，因此，connect()方法返回时，连接并不一定建立成功，在后续使用这个连接时，还需要使用finishConnect()再次确认。第7行，将这个Channel和Selector进行绑定，并注册了感兴趣的事件作为连接（OP_CONNECT）。

初始化完成后，就是程序的主要执行逻辑：

```
01 public void working() throws IOException {
02     while (true) {
03         if (!selector.isOpen())
04             break;
05         selector.select();
06         Iterator<SelectionKey> ite = this.selector.selectedKeys().iterator();
07         while (ite.hasNext()) {
08             SelectionKey key = ite.next();
09             ite.remove();
10             // 连接事件发生
11             if (key.isConnectable()) {
12                 connect(key);
13             } else if (key.isReadable()) {
14                 read(key);
15             }
16         }
17     }
}
```

在上述代码中，第5行通过Selector得到已经准备好的事件。如果当前没有任何事件准备就绪，这里就会阻塞。这里的整个处理机制和服务端非常类似，主要处理两个事件，首先是表示连接就绪的Connect事件（由connect()函数处理）以及表示通道可读的Read事件（由read()函数处理）。

函数connect()的实现如下：

```
01 public void connect(SelectionKey key) throws IOException {
02     SocketChannel channel = (SocketChannel) key.channel();
03     // 如果正在连接，则完成连接
04     if (channel.isConnectionPending()) {
05         channel.finishConnect();
06     }
07     channel.configureBlocking(false);
08     channel.write(ByteBuffer.wrap(new String("hello server!\r\n\
09         .getBytes())));
10     channel.register(this.selector, SelectionKey.OP_READ);
11 }
```

上述connect()函数接收SelectionKey作为其参数。在第4~6行，它首先判断是否连接已经建立，如果没有，则调用finishConnect()完成连接。建立连接后，向Channel写入数据，并同时注册读事件为感兴趣的事件（第10行）。

当Channel可读时，会执行read()方法，进行数据读取：

```
01 public void read(SelectionKey key) throws IOException {
02     SocketChannel channel = (SocketChannel) key.channel();
03     // 创建读取的缓冲区
04     ByteBuffer buffer = ByteBuffer.allocate(100);
05     channel.read(buffer);
06     byte[] data = buffer.array();
07     String msg = new String(data).trim();
08     System.out.println("客户端收到信息: " + msg);
09     channel.close();
10     key.selector().close();
11 }
```

上述read()函数首先创建了100字节的缓冲区（第4行），接着从Channel中读取数据，并将其打印在控制台上。最后，关闭Channel和Selector。

5.11 读完了再通知我：AIO

AIO是异步IO的缩写，即Asynchronized。虽然NIO在网络操作中，提供了非阻塞的方法，但是NIO的IO行为还是同步的。对于NIO来说，我们的业务线程是在IO操作准备好时，得到通知，接着就由这个线程自行进行IO操作，IO操作本身还是同步的。

但对于AIO来说，则更加进了一步，它不是在IO准备好时再通知线程，而是在IO操作已经完成后，再给线程发出通知。因此，AIO是完全不会阻塞的。此时，我们的业务逻辑将变成一个回调函数，等待IO操作完成后，由系统自动触发。

下面，我将通过AIO来实现一个简单的EchoServer以及对应的客户端。

5.11.1 AIO EchoServer的实现

异步IO需要使用异步通道（AsynchronousServerSocketChannel）：

```
public final static int PORT = 8000;
private AsynchronousServerSocketChannel server;
public AIOEchoServer() throws IOException {
    server = AsynchronousServerSocketChannel.open().bind(new Inet
}
```

上述代码绑定了8000端口为服务器端口，并使用

AsynchronousServerSocketChannel异步Channel作为服务器，变量名为server。

我们使用这个server来进行客户端的接收和处理：

```
01 public void start() throws InterruptedException, ExecutionExce
02     System.out.println("Server listen on " + PORT);
03     //注册事件和事件完成后的处理器
04     server.accept(null, new CompletionHandler<AsynchronousSocI
05         final ByteBuffer buffer = ByteBuffer.allocate(1024);
06         public void completed(AsynchronousSocketChannel result
07             System.out.println(Thread.currentThread().getName(
08             Future<Integer> writeResult=null;
09             try {
10                 buffer.clear();
11                 result.read(buffer).get(100, TimeUnit.SECONDS)
12                 buffer.flip();
13                 writeResult=result.write(buffer);
14             } catch (InterruptedException | ExecutionException
15                 e.printStackTrace();
16             } catch (TimeoutException e) {
17                 e.printStackTrace();
18             } finally {
19                 try {
20                     server.accept(null, this);
21                     writeResult.get();
22                     result.close();
```

```

23         } catch (Exception e) {
24             System.out.println(e.toString());
25         }
26     }
27 }
28
29 @Override
30 public void failed(Throwable exc, Object attachment) {
31     System.out.println("failed: " + exc);
32 }
33 });
34 }

```

上述定义的`start()`方法开启了服务器。值得注意的是，这个方法除了第2行的打印语句外，只调用了一个函数`server.accept()`。之后，你看到的那一大堆代码只是这个函数的参数。

`AsynchronousServerSocketChannel.accept()`方法会立即返回。它并不会真的去等待客户端的到来。在这里使用的`accept()`方法的签名为：

```

public final <A> void accept(A attachment,
                             CompletionHandler<AsynchronousSocketChannel,?

```

它的第一个参数是一个附件，可以是任意类型，作用是让当前线程和后续的回调方法可以共享信息，它会在后续调用中，传递给`handler`。它的第二个参数是`CompletionHandler`接口。这个接口有两个方法：

```

void completed(V result, A attachment);

```



```
void failed(Throwable exc, A attachment);
```

这两个方法分别在异步操作accept()成功或者失败时被回调。

因此AsynchronousServerSocketChannel.accept()实际上做了两件事，第一是发起accept请求，告诉系统可以开始监听端口了。第二，注册CompletionHandler实例，告诉系统，一旦有客户端前来连接，如果成功连接，就去执行CompletionHandler.completed()方法；如果连接失败，就去执行CompletionHandler.failed()方法。

所以，server.accept()方法不会阻塞，它会立即返回。

下面，来分析一下CompletionHandler.completed()的实现。当completed()被执行时，意味着已经有客户端成功连接了。在第11行，使用read()方法读取客户的数据。这里要注意，AsynchronousSocketChannel.read()方法也是异步的，换句话说它不会等待读取完成了再返回，而是立即返回，返回的结果是一个Future，因此这里就是Future模式的典型应用。为了编程方便，我在这里直接调用Future.get()方法，进行等待，将这个异步方法变成了同步方法。因此，在第11行执行完成后，数据读取就已经完成了。

之后，将数据回写给客户端（第13行）。这里调用的是AsynchronousSocketChannel.write()方法。这个方法不会等待数据全部写完，也是立即返回的。同样，它返回的也是Future对象。

再之后，在第20行，服务器进行下一个客户端连接的准备。同时关闭当前正在处理的客户端连接。但在关闭之前，得先确保之前的write()操作已经完成，因此，使用Future.get()方法进行等待（第21行）。

接下来，我们只需要在主函数中调用这个`start()`方法就可以开启服务器了：

```
01 public static void main(String args[]) throws Exception {
02     new AIOEchoServer().start();
03     // 主线程可以继续自己的行为
04     while (true) {
05         Thread.sleep(1000);
06     }
07 }
```

上述代码第2行，调用`start()`方法开启服务器。但由于`start()`方法里使用的都是异步方法，因此它会马上返回，它并不像阻塞方法那样会进行等待。因此，如果想让程序驻守执行，第4~6行的等待语句是必需的。否则，在`start()`方法结束后，不等客户端到来，程序已经运行完成，主线程就将退出。

5.11.2 AIO Echo客户端实现

在服务端的实现中，我们使用`Future.get()`方法将异步调用转为了一个同步等待。在客户端的实现里，我们将全部使用异步回调实现：

```
01 public class AIOClient {
02     public static void main(String[] args) throws Exception {
03         final AsynchronousSocketChannel client = AsynchronousS
04         client.connect(new InetSocketAddress("localhost", 8000), nul
05             @Override
```

```

06         public void completed(Void result, Object attachme
07     client.write(ByteBuffer.wrap("Hello!".getBytes()), null, ne
08         @Override
09         public void completed(Integer result, Obje
10             try {
11                 ByteBuffer buffer = ByteBuffer.all
12     client.read(buffer,buffer,new CompletionHa
13         @Override
14         public void completed(Integer
15             buffer.flip();
16             System.out.println(new Str
17             try {
18                 client.close();
19             } catch (IOException e) {
20                 e.printStackTrace();
21             }
22         }
23         @Override
24         public void failed(Throwable e
25         }
26         });
27     } catch (Exception e) {
28         e.printStackTrace();
29     }
30 }
31 @Override
32 public void failed(Throwable exc, Object a

```

```

33         }
34     });
35 }
36 @Override
37     public void failed(Throwable exc, Object attachmen
38     }
39 });
40     //由于主线程马上结束，这里等待上述处理全部完成
41     Thread.sleep(1000);
42 }
43 }

```

上面的AIO客户端看起来代码很长，但实际上只有三个语句。

第一个语句为第3行，打开AsynchronousSocketChannel通道。第二个语句是第4~39行，它让客户端去连接指定的服务器，并注册了一系列事件。第三个语句是第41行，让线程进行等待。虽然第2个语句看起来很长，但是它完全是异步的，因此会很快返回，并不会等待在连接操作的过程中。如果不进行等待，客户端会马上退出，也就无法继续工作了。

第4行，客户端进行网络连接，并注册了连接成功的回调函数CompletionHandler<Void, Object>。待连接成功后，就会进入代码第7行。第7行进行数据写入，向服务端发送数据。这个过程也是异步的，会很快返回。写入完成后，会通知回调接口CompletionHandler<Integer, Object>，进入第10行。第10行开始，准备进行数据读取，从服务端读取回写的的数据。当然，第12行的read()函数也是立即返回的，成功读取所有数据后，会回调CompletionHandler<Integer, ByteBuffer>接口，进

入第15行。在第15～16行，打印接收到的数据。

5.12 参考文献

- 有关disruptor的性能测试
 - <https://github.com/LMAX-Exchange/disruptor/wiki/Performance-Results>
- disruptor的小例子
 - <https://github.com/LMAX-Exchange/disruptor/wiki/Getting-Started>
- 伪共享的案例
 - <http://mechanical-sympathy.blogspot.jp/2011/07/false-sharing.html>
- 有关并行排序的更详细资料
 - The Art of Concurrency: A Thread Monkey's Guide to Writing Parallel Applications
 - 并发的艺术Clay Brebears著
- 一篇描述NIO网络编程的博客
 - <http://weixiaolu.iteye.com/blog/1479656>
- 有关NIO Buffer的使用
 - http://www.uucode.net/201504/nio_buffer

- http://www.uucode.net/201504/buffer_operations
- http://www.uucode.net/201504/nio_buffer_channel
- 有关AIO的一篇入门帖子
 - <http://www.iteye.com/topic/1113611>

第6章 Java 8与并发

2014年，Oracle发布了Java 8新版本。对于Java来说，这显然是一个具有里程碑意义的版本。它最主要的改进是增加了函数式编程的功能。就目前来说，Java最令人头痛的问题，也是受到最多质疑的地方，应该就是Java那烦琐的语法。这样我们不得不花费大量的代码行数，来实现一些司空见惯的功能，以至于Java程序总是冗长的。但是，这一切将在Java 8的函数式编程中得到缓解。

严格来说，函数式编程与我们的主题并没有太大关系，我似乎不应该在这里提及它。但是，在Java 8中新增的一些与并行相关的API，却以函数式编程的范式出现，为了能让大家更好地理解这些功能，我会先简要地介绍一下Java 8中的函数式编程。

6.1 Java 8的函数式编程简介

函数式编程与面向对象的设计方法在思路 and 手段上都各有千秋，在这里，我将简要介绍一下函数式编程与面向对象相比较的一些特点和差异。

6.1.1 函数作为一等公民

在理解函数作为一等公民这句话时，让我们先来看一下一种非常常用的互联网语言JavaScript，相信大家对它都不会陌生。JavaScript并不是严格意义上的函数式编程，不过，它也不是属于严格的面向对象。但是，如果你愿意，你既可以把它当作面向对象语言，也可以把它当作函数式语言，因此，称之为多范式语言，可能更加合适。

如果你使用jQuery，你可能会经常使用如下的代码：

```
$("#button").click(function(){
    $("#li").each(function(){
        alert($(this).text())
    });
});
```

注意这里each()函数的参数，这是一个匿名函数，在遍历所有的li节点时，会弹出li节点的文本内容。将函数作为参数传递给另外一个函数，这是函数式编程的特性之一。

再来考察另外一个案例：

```
function f1(){
    var n=1;
    function f2(){
        alert(n);
    }
    return f2;
}
var result=f1();
result(); // 1
```

这也是一段JavaScript代码。在这段代码中，注意函数f1的返回值，它返回了函数f2。在倒数第2行，返回的f2函数并赋值给result，实际上，此时的result就是一个函数，并且指向f2。对result的调用，就会打印n的值。

函数可以作为另外一个函数的返回值，也是函数式编程的重要特点。

6.1.2 无副作用

函数的副作用指的是函数在调用过程中，除了给出了返回值外，还修改了函数外部的状态。比如，函数在调用过程中，修改了某一个全局状态。函数式编程认为，函数的副作用应该被尽量避免。可以想象，如果一个函数肆意修改全局或者外部状态，当系统出现问题时，我们可能很难判断究竟是哪个函数引起的问题，这对于程序的调试和跟踪是没有

好处的。如果函数都是显式函数，那么函数的执行显然不会受到外部或者全局信息的影响，因此，对于调试和排错是有益的。

注意：显式函数指函数与外界交换数据的唯一渠道就是参数和返回值，显式函数不会去读取或者修改函数的外部状态。与之相对的是隐式函数，隐式函数除了参数和返回值外，还会读取外部信息，或者可能修改外部信息。

然而，完全的无副作用实际上做不到的，因为系统总是需要获取或者修改外部信息的，同时，模块之间的交互也极有可能是通过共享变量进行的。如果完全禁止副作用的出现，也是一件让人很不愉快的事情。因此，大部分函数式编程语言，如Clojure等，都允许副作用的存在。但是与面向对象相比，这种函数调用的副作用，在函数式编程里，需要进行有效的限制。

6.1.3 申明式的（**Declarative**）

函数式编程是申明式的编程方式。相对于命令式（**Imperative**）而言，命令式的程序设计喜欢大量使用可变对象和指令。我们总是习惯于创建对象或者变量，并且修改它们的状态或者值，或者喜欢提供一系列指令，要求程序执行。这种编程习惯在申明式的函数式编程中有所变化。对于申明式的编程范式，你不再需要提供明确的指令操作，所有的细节指令将会更好地被程序库所封装，你要做的只是提出你的要求，申明你的用意即可。

请看下面一段程序，这一段传统的命令式编程，为了打印数组中的值，我们需要进行一个循环，并且每次需要判断循环是否结束。在循环

体内，我们要明确地给出需要执行的语句和参数。

```
public static void imperative(){
    int[] iArr={1,3,4,5,6,9,8,7,4,2};
    for(int i=0;i<iArr.length;i++){
        System.out.println(iArr[i]);
    }
}
```

与之对应的申明式代码如下：

```
public static void declarative(){
    int[] iArr={1,3,4,5,6,9,8,7,4,2};
    Arrays.stream(iArr).forEach(System.out::println);
}
```

可以看到，变量数组的循环体居然消失了！println()函数似乎在这里也没有指定任何参数，在此，我们只是简单地申明了我们的用意。有关循环以及判断循环是否结束等操作都被简单地封装在程序库中。

6.1.4 不变的对象

在函数式编程中，几乎所有传递的对象都不会被轻易修改。

请看以下代码：

```
static int[] arr={1,3,4,5,6,7,8,9,10};
Arrays.stream(arr).map((x)->x=x+1).forEach(System.out::println);
```

```
System.out.println();  
Arrays.stream(arr).forEach(System.out::println);
```

代码第2行看似对每一个数组成员执行了加1的操作。但是在操作完成后，在最后一行，打印`arr`数组所有的成员值时，你还是会发现，数组成员并没有变化！在使用函数式编程时，这种状态是一种常态，几乎所有的对象都拒绝被修改。这非常类似于不变模式。

6.1.5 易于并行

由于对象都处于不变的状态，因此函数式编程更加易于并行。实际上，你甚至完全不用担心线程安全的问题。我们之所以要关注线程安全，一个很重要的原因是当多个线程对同一个对象进行写操作时，容易将这个对象“写坏”。但是，由于对象是不变的，因此，在多线程环境下，也就没有必要进行任何同步操作。这样不仅有利于并行化，同时，在并行化后，由于没有同步和锁机制，其性能也会比较好。

6.1.6 更少的代码

通常情况下，函数式编程更加简明扼要，Clojure语言（一种运行于JVM的函数式语言）的爱好者就宣称，使用Clojure可以将Java代码行数减少到原有的十分之一。一般说来，精简的代码更易于维护。引入函数式编程范式后，我们可以使用Java用更少的代码完成更多的工作。

请看下面这个例子，对于数组中每一个成员，首先判断是否是奇数，如果是奇数，则执行加1，并最终打印数组内所有成员。

数组定义：

```
static int[] arr={1,3,4,5,6,7,8,9,10};
```

传统的处理方式：

```
for(int i=0;i<arr.length;i++){  
    if(arr[i]%2!=0){  
        arr[i]++;  
    }  
    System.out.println(arr[i]);  
}
```

使用函数式方式：

```
Arrays.stream(arr).map(x->(x%2==0?x:x+1)).forEach(System.out::pr:
```

可以看到，函数式范式更加紧凑而且简洁。

6.2 函数式编程基础

在正式进入函数式编程之前，有必要先了解一下Java 8为支持函数式编程所做的基础性的改进，这里，将简要介绍一下FunctionalInterface注释、接口默认方法和方法句柄。

6.2.1 FunctionalInterface注释

Java 8提出了函数式接口的概念。所谓函数式接口，简单来说，就是只定义了单一抽象方法的接口。比如下面的定义：

```
@FunctionalInterface
public static interface IntHandler{
    void handle(int i);
}
```

注释FunctionalInterface用于表明IntHandler接口是一个函数式接口，该接口被定义为只包含一个抽象方法handle()，因此它符合函数式接口的定义。如果一个函数满足函数式接口的定义，那么即使不标注为@FunctionalInterface，编译器依然会把它看做函数式接口。这有点像@Override注释，如果你的函数符合重载的要求，无论你是否标注了@Override，编译器都会识别这个重载函数，但一旦你进行了标注，而实际的代码不符合规范，那么就会得到一个编译错误。如图6.1所示，展示了一个不符合规范，却被标注为@FunctionalInterface的接口。很显然，该IntHandler包含两个抽象方法，因此不符合函数式接口的要求，

又因为IntHandler接口被标注为函数式接口，产生矛盾，故编译出错。

```
9  @FunctionalInterface
10 public static interface IntHandler{
11     void handle(int i);
12     void handle2(int i);
13 }
```

图6-1 不符合规范的函数式接口

这里需要强调的是，函数式接口只能有一个抽象方法，而不是只能有一个方法。这分两点来说明：首先，在Java 8中，接口运行存在实例方法（参见下节的“接口默认方法”），其次任何被java.lang.Object实现的方法，都不能视为抽象方法，因此，下面的NonFunc接口不是函数式接口，因为equals()方法在java.lang.Object中已经实现。

```
interface NonFunc {
    boolean equals(Object obj);
}
```

同理，下面实现的IntHandler接口符合函数式接口要求，虽然看起来它不像，但实际上它是一个完全符合规范的函数式接口。

```
@FunctionalInterface
public static interface IntHandler{
    void handle(int i);
    boolean equals(Object obj);
}
```

函数式接口的实例可以由方法引用或者lambda表达式进行构造，这个我们将在后面进一步举例说明。

6.2.2 接口默认方法

在Java 8之前的版本，接口只能包含抽象方法。但从Java 8之后，接口也可以包含若干个实例方法。这一改进使得Java 8拥有了类似于多继承的能力。一个对象实例，将拥有来自于多个不同接口的实例方法。

比如，对于接口IHorse，实现如下：

```
public interface IHorse{  
    void eat();  
    default void run(){  
        System.out.println("hourse run");  
    }  
}
```

在Java 8中，使用default关键字，可以在接口内定义实例方法。注意，这个方法并非抽象方法，而是拥有特定逻辑的具体实例方法。

所有的动物都能自由呼吸，所以，这里可以再定义一个IAntimal接口，它也包含一个默认方法breath()。

```
public interface IAnimal {  
    default void breath(){  
        System.out.println("breath");  
    }  
}
```

骡是马和驴的杂交物种，因此骡（Mule）可以实现为IHorse，同时

骡也是动物，因此有：

```
public class Mule implements IHorse, IAnimal{
    @Override
    public void eat() {
        System.out.println("Mule eat");
    }
    public static void main(String[] args) {
        Mule m=new Mule();
        m.run();
        m.breath();
    }
}
```

注意上述代码中Mule实例同时拥有来自不同接口的实现方法。这在Java 8之前是做不到的。从某种程度上说，这种模式可以弥补Java单一继承的一些不便。但同时也要知道，它也将遇到和多继承相同的问题，如图6.2所示。如果IDonkey也存在一个默认的run()方法，那么同时实现它们的Mule，就会不知所措，因为它不知道应该以哪个方法为准。

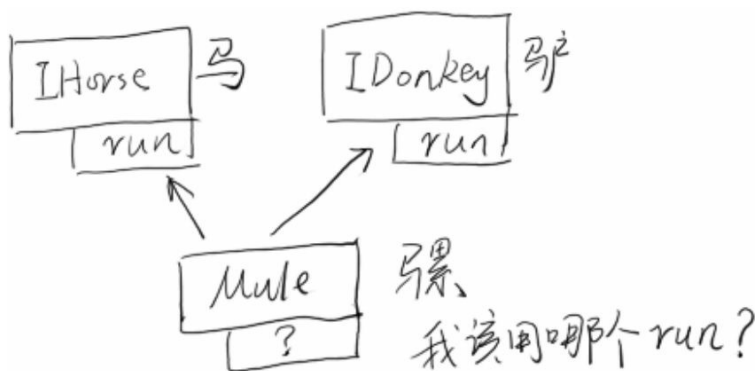


图6-2 接口默认方法带来的多继承问题

增加一个IDonkey的实现:

```
public interface IDonkey{
    void eat();
    default void run(){
        System.out.println("Donkey run");
    }
}
```

修改骡Mule的实现如下, 注意它同时实现了IHorse和IDonkey:

```
public class Mule implements IHorse, IDonkey, IAnimal{
    @Override
    public void eat() {
        System.out.println("Mule eat");
    }
    public static void main(String[] args) {
        Mule m=new Mule();
        m.run();
        m.breath();
    }
}
```

此时, 由于IHorse和IDonkey拥有相同的默认实例方法, 故编译器会抛出一个错误:

```
Duplicate default methods named run with the parameters () and ()
```

IDonkey and IHorse

为了让Mule同时实现IHorse和IDonkey，在这里，我们不得不重新实现一下run()方法，让编译器可以进行方法绑定。修改Mule的实现如下：

```
public class Mule implements IHorse, IDonkey, IAnimal{

    @Override
    public void run(){
        IHorse.super.run();
    }

    @Override
    public void eat() {
        System.out.println("Mule eat");
    }

    public static void main(String[] args) {
        Mule m=new Mule();
        m.run();
        m.breath();
    }
}
```

在这里，将Mule的run()方法委托给IHorse实现，当然，大家也可以有自己的实现。

接口默认实现对于整个函数式编程的流式表达非常重要。比如，大家熟悉的java.util.Comparator接口，它在JDK 1.2时就已经被引入，用于

在排序时给出两个对象实例的具体比较逻辑。在Java 8中，Comparator接口新增了若干个默认方法，用于多个比较器的整合。其中一个常用的默认方法如下：

```
default Comparator<T> thenComparing(Comparator<? super T> other) {
    Objects.requireNonNull(other);
    return (Comparator<T> & Serializable) (c1, c2) -> {
        int res = compare(c1, c2);
        return (res != 0) ? res : other.compare(c1, c2);
    };
}
```

有了这个默认方法，在进行排序时，我们就可以非常方便地进行元素的多条件排序，比如，如下代码构造一个比较器，它先按照字符串长度排序，继而按照大小写不敏感的字母顺序排序。

```
Comparator<String> cmp = Comparator.comparingInt(String::length)
    .thenComparing(String.CASE_INSENSITIVE_ORDER);
```

6.2.3 lambda表达式

lambda表达式可以说是函数式编程的核心。lambda表达式即匿名函数，它是一段没有函数名的函数体，可以作为参数直接传递给相关的调用者。lambda表达式极大地增强了Java语言的表达能力。

下例展示了lambda表达式的使用，在forEach()函数中，传入的就是一个lambda表达式，它完成了对元素的标准输出操作。可以看到这段表

达式并不像函数一样有名字，非常类似匿名内部类，它只是简单地描述了应该执行的代码段。

```
List<Integer> numbers = Arrays.asList(1, 2, 3, 4, 5, 6);  
numbers.forEach((Integer value) -> System.out.println(value));
```

和匿名对象一样，lambda表达式也可以访问外部的局部变量，如下所示：

```
final int num = 2;  
Function<Integer, Integer> stringConverter = (from) -> from * num;  
System.out.println(stringConverter.apply(3));
```

上述代码可以编译通过，正常执行，并输出6。与匿名内部对象一样，在这种情况下，外部的num变量必须申明为final，这样才能保证在lambda表达式中合法的访问它。

但奇妙的是，对于lambda表达式而言，即使去掉上述的final定义，程序依然可以编译通过！但千万不要以为这样你就可以修改num的值了。实际上，这只是Java 8做了一个掩人耳目的小处理，它会自动地将在lambda表达式中使用的变量视为final。因此，下述代码是可以编译通过的：

```
int num = 2;  
Function<Integer, Integer> stringConverter = (from) -> from * num;  
System.out.println(stringConverter.apply(3));
```

但是，如果像下面这么写，就不行：

```
int num = 2;
```

```
Function<Integer, Integer> stringConverter = (from) -> from * n
    num++;
System.out.println(stringConverter.apply(3));
```

上述的num++会引起一个编译错误：

```
Local variable num defined in an enclosing scope must be final or
```

6.2.4 方法引用

方法引用是Java 8中提出的用来简化lambda表达式的一种手段。它通过类名和方法名来定位到一个静态方法或者实例方法。

方法引用在Java 8中的使用非常灵活。总的来说，可以分为以下几种。

- 静态方法引用：ClassName::methodName
- 实例上的实例方法引用：instanceReference::methodName
- 超类上的实例方法引用：super::methodName
- 类型上的实例方法引用：ClassName::methodName
- 构造方法引用：Class::new
- 数组构造方法引用：TypeName[]::new

首先，方法引用使用“::”定义，“::”的前半部分表示类名或者实例名，后半部分表示方法名称。如果是构造函数，则使用new表示。

下例展示了方法引用的基本使用：

```

public class InstanceMethodRef {
    public static void main(String[] args) {
        List<User> users=new ArrayList<User>();
        for(int i=1;i<10;i++){
            users.add(new User(i,"billy"+Integer.toString(i)));
        }
        users.stream().map(User::getName).forEach(System.out::pri
    }
}

```

对于第1个方法引用“**User::getName**”，表示User类的实例方法。在执行时，Java会自动识别流中的元素（这里指User实例）是作为调用目标还是调用方法的参数。在“**User::getName**”中，显然流内的元素都应该作为调用目标，因此实际上，在这里调用了每一个User对象实例的getName()方法，并将这些User的name作为一个新的流。同时，对于这里得到的所有name，使用方法引用System.out::println进行处理。这里的System.out为PrintStream对象实例，因此，这里表示System.out实例的println方法，系统也会自动判断，流内的元素此时应该作为方法的参数传入，而不是调用目标。

一般来说，如果使用的是静态方法，或者调用目标明确，那么流内的元素会自动作为参数使用。如果函数引用表示实例方法，并且不存在调用目标，那么流内元素就会自动作为调用目标。

因此，如果一个类中存在同名的实例方法和静态函数，那么编译器就会感到很困惑，因为此时，它不知道应该使用哪个方法进行调用。它既可以选择同名的实例方法，将流内元素作为调用目标，也可以使用静态方法，将流元素作为参数。

请看下面的例子：

```
public class BadMethodRef {  
    public static void main(String[] args) {  
        List<Double> numbers=new ArrayList<Double>();  
        for(int i=1;i<10;i++){  
            numbers.add(Double.valueOf(i));  
        }  
        numbers.stream().map(Double::toString).forEach(System.out  
    }  
}
```

上述代码试图将所有的Double元素转为String并将其输出，但是很不幸，在Double中同时存在以下两个函数：

```
public static String toString(double d)  
public String toString()
```

此时，对函数引用的处理就出现了歧义，因此，这段代码在编译时就会抛出如下错误：

```
Ambiguous method reference: both toString() and toString(double)  
eligible
```

方法引用也可以使用构造函数。首先，查看模型类User的定义：

```
public class User{  
    private int id;
```

```

private String name;

public User(int id,String name){
    this.id=id;
    this.name=name;
}
//这里省略对字段的setter和getter
}

```

下面的方法引用调用了User的构造函数：

```

public class ConstrMethodRef {
    @FunctionalInterface
    interface UserFactory<U extends User> {
        U create(int id, String name);
    }

    static UserFactory<User> uf=User::new;

    public static void main(String[] args) {
        List<User> users=new ArrayList<User>();
        for(int i=1;i<10;i++){
            users.add(uf.create(i, "billy"+Integer.toString(i)));
        }
        users.stream().map(User::getName).forEach(System.out::pri
    }
}

```

在此，UserFactory作为User的工厂类，是一个函数式接口。当使用User::new创建接口实例时，系统会根据UserFactory.create()的函数签名来选择合适的User构造函数，在这里，很显然就是public User(int id,String name)。在创建UserFactory实例后，对UserFactory.create()的调用，都会委托给User的实际构造函数进行，从而创建User对象实例。

6.3 一步一步走入函数式编程

在了解了Java 8的一些新特性后，就可以正式开始进入函数式编程了。为了能让大家更快地理解函数式编程，我们先从简单的例子开始。

```
static int[] arr={1,3,4,5,6,7,8,9,10};

public static void main(String[] args) {
    for(int i:arr){
        System.out.println(i);
    }
}
```

上述代码循环遍历了数组内的元素，并且进行了数值的打印，这也是传统的做法。如果使用Java 8中的流，那么可以写成这样：

```
static int[] arr = { 1, 3, 4, 5, 6, 7, 8, 9, 10 };

public static void main(String[] args) {
    Arrays.stream(arr).forEach(new IntConsumer() {
        @Override
        public void accept(int value) {
            System.out.println(value);
        }
    });
}
```

注意：`Arrays.stream()`方法返回了一个流对象。类似于集合或者数组，流对象也是一个对象的集合，它将给予我们遍历处理流内元素的功能。

这里值得注意的是这个流对象的`forEach()`方法，它接收一个`IntConsumer`接口的实现，用于对每个流内的对象进行处理。之所以是`IntConsumer`接口，因为当前流是`IntStream`，也就是装有`Integer`元素的流，因此，它自然需要一个处理`Integer`元素的接口。函数`forEach()`会挨个将流内的元素送入`IntConsumer`进行处理，循环过程被封装在`forEach()`内部，也就是JDK框架内。

除了`IntStream`流外，`Arrays.stream()`还支持`DoubleStream`、`LongStream`和普通的对象流`Stream`，这完全取决于它所接受的参数，如图6.3所示。

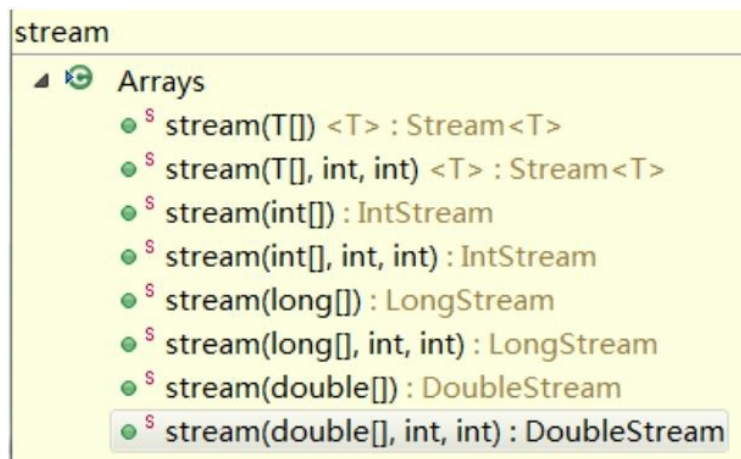


图6.3 Stream流的几种类型

但这样的写法可能还不能让人满意，代码量似乎比原先更多，而且除了引入了不必要的接口和匿名类等复杂性外，似乎也看不出来有什么太大的好处。但是，我们的脚步并未就此打住。试想，既然`forEach()`函数的参数是可以从上下文中推导出来的，那为什么还要不厌其烦地写出

来呢？这些机械的推导工作，就交给编译器去做吧！于是：

```
static int[] arr={1,3,4,5,6,7,8,9,10};

public static void main(String[] args) {
    Arrays.stream(arr).forEach((final int x)-> {
        System.out.println(x);
    });
}
```

从上述代码中可以看到，`IntStream`接口名称被省略了，这里只使用了参数名和一个实现体，看起来简洁很多了。但是还不够，因为参数的类型也是可以推导的。既然是`IntConsumer`接口，参数自然是`int`了，于是：

```
static int[] arr={1,3,4,5,6,7,8,9,10};

public static void main(String[] args) {
    Arrays.stream(arr).forEach((x)-> {
        System.out.println(x);
    });
}
```

好了，现在连参数类型也省略了，但是这两个花括号特别碍眼。虽然它们对程序没有什么影响，但是为了简单的一句执行语句要加上一对花括号也实属没有必要，那干脆也去掉吧！去掉花括号后，为了清晰起见，把参数申明和接口实现就放在一行吧！

```
static int[] arr={1,3,4,5,6,7,8,9,10};

public static void main(String[] args) {
    Arrays.stream(arr).forEach((x)->System.out.println(x));
}
```

这样看起来就好多了。此时，forEach()函数的参数依然是IntConsumer，但是它却以一种新的形式被定义，这就是lambda表达式。表达式由“->”分割，左半部分表示参数，右半部分表示实现体。因此，我们也可以简单地理解lambda表达式只是匿名对象实现的一种新的方式。实际上，也是这样的。

有兴趣的读者可以使用虚拟机参数-Djdk.internal.lambda.dumpProxyClasses启动带有lambda表达式的Java小程序，该参数会将lambda表达式相关的中间类型进行输出，方便调试和学习。在本例中，输出了HelloFunction6\$\$Lambda\$1.class类，使用以下命令进行并发汇编操作：

```
javap -p -v HelloFunction6$$Lambda$1.class
```

在输出结果中，可以清楚地看到：

```
final class geym.java8.func.ch3.HelloFunction6$$Lambda$1 implements
java.util.function.IntConsumer
```

省略部分输出

```
public void accept(int);
    descriptor: (I)V
    flags: ACC_PUBLIC
```

Code:

```
stack=1, locals=2, args_size=2
0: iload_1
1: invokestatic #17 // Method geym/java8/func/ch3/HelloFun
4: return
```

限于篇幅有限，这里只给出了我们关心的内容。首先，这个中间类型确实实现了IntConsumer接口。其次，在实现accept()方法时，它内部委托给了一个名为HelloFunction6.lambda\$0()的方法。可以推测，这个方法也是编译时自动生成的。

使用以下命令查看HelloFunction6的编译结果：

```
javap -p -v HelloFunction6
```

我们很惊喜地找到了期待已久的lambda\$0()方法，其实现如下：

```
private static void lambda$0(int);
descriptor: (I)V
flags: ACC_PRIVATE, ACC_STATIC, ACC_SYNTHETIC
Code:
    stack=2, locals=1, args_size=1
        0: getstatic      #41                // Field java/lang/Syst
        3: iload_0
        4: invokevirtual #47                // Method java/io/Print
        7: return
```

它被实现为一个私有的静态方法，实现内容就是简单地进行了System.out.println()的调用，也正是我们代码中lambda表达式的内容。

由此，可以看到，Java 8中对lambda表达式的处理几乎等同于匿名类的实现，但是在写法上和编程范式上有了明显的区别。

不过，简化代码的流程并没有结束，在上一节中已经提到，Java 8还支持了方法引用，通过方法引用的推导，你甚至连参数申明和传递都可以省略。

```
static int[] arr={1,3,4,5,6,7,8,9,10};

public static void main(String[] args) {
    Arrays.stream(arr).forEach(System.out::println);
}
```

至此，欢迎大家正式进入Java 8函数式编程的殿堂，那些看似玄妙的lambda表达式的解析和工作原理已经介绍完毕。

使用lambda表达式不仅可以简化匿名类的编写，与接口的默认方法相结合，还可以使用更顺畅的流式API对各种组件进行更自由的装配。

下面这个例子对集合中所有元素进行两次输出，一次输出到标准错误，一次输出到标准输出中。

```
static int[] arr={1,3,4,5,6,7,8,9,10};

public static void main(String[] args) {
    IntConsumer outprintln=System.out::println;
    IntConsumer errprintln=System.err::println;
    Arrays.stream(arr).forEach(outprintln.andThen(errprintln));
}
```

这里首先使用函数引用，直接定义了两个IntConsumer接口实例，一个指向标准输出，另一个指向标准错误。使用接口默认函数IntConsumer.addThen()，将两个IntConsumer进行组合，得到一个新的IntConsumer，这个新的IntConsumer会依次调用outprintln和errprintln，完成对数组中元素的处理。

其中IntConsumer.addThen()的实现如下，仅供大家参考：

```
default IntConsumer andThen(IntConsumer after) {  
    Objects.requireNonNull(after);  
    return (int t) -> { accept(t); after.accept(t); };  
}
```

可以看到，addThen()方法返回一个新的IntConsumer，这个新的IntConsumer会先调用第1个IntConsumer进行处理，接着调用第2个IntConsumer处理，从而实现多个处理器的整合。这种操作手法在Java 8的函数式编程中极其常见，请大家留意。

6.4 并行流与并行排序

Java 8中，可以在接口不变的情况下，将流改为并行流。这样，就可以很自然地使用多线程进行集合中的数据处理。

6.4.1 使用并行流过滤数据

现在让我们考虑这么一个简单的案例，我们希望可以统计1~1000000内所有的质数的数量。首先，我们需要一个判断质数的函数：

```
public class PrimeUtil {  
    public static boolean isPrime(int number) {  
        int tmp = number;  
        if (tmp < 2) {  
            return false;  
        }  
        for (int i = 2; Math.sqrt(tmp) >= i; i++) {  
            if (tmp % i == 0) {  
                return false;  
            }  
        }  
        return true;  
    }  
}
```

上述函数给定一个数字，如果这个数字是质数就返回true，否则返回false。

接着，使用函数式编程统计给定范围内所有的质数：

```
IntStream.range(1, 1000000).filter(PrimeUtil::isPrime).count();
```

上述代码首先生成一个1到1000000的数字流。接着使用过滤函数，只选择所有的质数，最后进行数量统计。

上述代码是串行的，将它改造成并行计算非常简单，只需要将流并行化即可：

```
IntStream.range(1, 1000000).parallel().filter(PrimeUtil::isPrime)
```

上述代码中，首先parallel()方法得到一个并行流，接着，在并行流上进行过滤，此时，PrimeUtil.isPrime()函数会被多线程并发调用，应用于流中的所有元素。

6.4.2 从集合得到并行流

在函数式编程中，我们可以从集合得到一个流或者并行流。下面这段代码试图统计集合内所有学生的平均分：

```
List<Student> ss=new ArrayList<Student>();  
double ave=ss.stream().mapToInt(s->s.score).average().getAsDouble();
```

从集合对象List中，我们使用stream()方法可以得到一个流。如果希望将这段代码并行化，则可以使用parallelStream()函数。

```
double ave=ss.parallelStream().mapToInt(s->s.score).average().get
```

可以看到，将原有的串行方式改造成并行执行是非常容易的。

6.4.3 并行排序

除了并行流外，对于普通数组，Java 8中也提供了简单的并行功能。比如，对于数组排序，我们有Arrays.sort()方法。当然这是串行排序，但在Java 8中，我们可以使用新增的Arrays.parallelSort()方法直接使用并行排序。

比如，你可以这样使用：

```
int[] arr=new int[10000000];  
Arrays.parallelSort(arr);
```

除了并行排序外，Arrays中还增加了一些API用于数组中数据的赋值，比如：

```
public static void setAll(int[] array, IntUnaryOperator generator
```

这是一个函数式味道很浓的接口，它的第2个参数是一个函数式接口。如果我们想给数组中每一个元素都附上一个随机值，则可以这么做：

```
Random r=new Random();  
Arrays.setAll(arr, (i)->r.nextInt());
```

当然，以上过程是串行的。但是只要使用setAll()对应的并行版本，

你就可以很快将它执行在多个CPU上:

```
Random r=new Random();  
Arrays.parallelSetAll (arr, (i)->r.nextInt());
```

6.5 增强的Future: CompletableFuture

CompletableFuture是Java 8新增的一个超大型工具类。为什么说它大呢？因为一方面，它实现了Future接口，而更重要的是，它也实现了CompletionStage接口。CompletionStage接口也是在Java 8中新增的。而CompletionStage接口拥有多达约40种方法！是的，你没有看错，这看起来完全不符合设计原则中所谓的“单方法接口”，但是在这里，它就这么存在了。这个接口之所以拥有如此众多的方法，是为了函数式编程中的流式调用准备的。通过CompletionStage提供的接口，我们可以在一个执行结果上进行多次流式调用，以此可以得到最终结果。比如，你可以在一个CompletionStage上进行如下调用：

```
stage.thenApply(x -> square(x)).thenAccept(x -> System.out.print  
System.out.println())
```

这一连串的调用就会挨个执行。

6.5.1 完成了就通知我

CompletableFuture和Future一样，可以作为函数调用的契约。如果你向CompletableFuture请求一个数据，如果数据还没有准备好，请求线程就会等待。而让人惊喜的是，通过CompletableFuture，我们可以手动设置CompletableFuture的完成状态。

```
01 public static class AskThread implements Runnable {
02     CompletableFuture<Integer> re = null;
03
04     public AskThread(CompletableFuture<Integer> re) {
05         this.re = re;
06     }
07
08     @Override
09     public void run() {
10         int myRe = 0;
11         try {
12             myRe = re.get() * re.get();
13         } catch (Exception e) {
14         }
15         System.out.println(myRe);
16     }
17 }
18
19 public static void main(String[] args) throws InterruptedException
20     final CompletableFuture<Integer> future = new CompletableFuture
21     new Thread(new AskThread(future)).start();
22     // 模拟长时间的计算过程
23     Thread.sleep(1000);
24     // 告知完成结果
25     future.complete(60);
26 }
```


上述代码在第1~17行，定义了一个AskThread线程。它接收一个CompletableFuture作为其构造函数，它的任务是计算CompletableFuture表示的数字的平方，并将其打印。

代码第20行，我们创建一个CompletableFuture对象实例，第21行，我们将这个对象实例传递给这个AskThread线程，并启动这个线程。此时，AskThread在执行到第12行代码时会阻塞，因为CompletableFuture中根本没有它所需要的数据，整个CompletableFuture处于未完成状态。第23行用于模拟长时间的计算过程。当计算完成后，可以将最终数据载入CompletableFuture，并标记为完成状态（第25行）。

当第25行代码执行后，表示CompletableFuture已经完成，因此AskThread就可以继续执行了。

6.5.2 异步执行任务

通过CompletableFuture提供的进一步封装，我们很容易实现Future模式那样的异步调用。比如：

```
01 public static Integer calc(Integer para) {
02     try {
03         // 模拟一个长时间的执行
04         Thread.sleep(1000);
05     } catch (InterruptedException e) {
06     }
07     return para*para;
08 }
```

```

09
10 public static void main(String[] args) throws InterruptedException
11 {
12     final CompletableFuture<Integer> future =
13         CompletableFuture.supplyAsync(() -> calc(50));
14     System.out.println(future.get());
15 }

```

上述代码中，第11~12行使用`CompletableFuture.supplyAsync()`方法构造一个`CompletableFuture`实例，在`supplyAsync()`函数中，它会在一个新的线程中，执行传入的参数。在这里，它会执行`calc()`方法。而`calc()`方法的执行可能是比较慢的，但是这不影响`CompletableFuture`实例的构造速度，因此`supplyAsync()`会立即返回，它返回的`CompletableFuture`对象实例就可以作为这次调用的契约，在将来任何场合，用于获得最终的计算结果。代码第13行，试图获得`calc()`的计算结果，如果当前计算没有完成，则调用`get()`方法的线程就会等待。

在`CompletableFuture`中，类似的工厂方法有以下几个：

```

static <U> CompletableFuture<U> supplyAsync(Supplier<U> suppl
static <U> CompletableFuture<U> supplyAsync(Supplier<U> suppl
static CompletableFuture<Void> runAsync(Runnable runnable);
static CompletableFuture<Void> runAsync(Runnable runnable, Execu

```

其中`supplyAsync()`方法用于那些需要有返回值的场景，比如计算某个数据等。而`runAsync()`方法用于没有返回值的场景，比如，仅仅是简单地执行某一个异步动作。

在这两对方法中，都有一个方法可以接收一个Executor参数。这就使我们可以让Supplier <U>或者Runnable在指定的线程池中工作。如果不指定，则在默认的系统公共的ForkJoinPool.common线程池中执行。

注意：在Java 8中，新增了ForkJoinPool.commonPool()方法。它可以获得一个公共的ForkJoin线程池。这个公共线程池中的所有线程都是Daemon线程。这意味着如果主线程退出，这些线程无论是否执行完毕，都会退出系统。

6.5.3 流式调用

在前文中我已经简单的提到，CompletionStage的约40个接口是为函数式编程做准备的。在这里，就让我们看一下，如何使用这些接口进行函数式的流式API调用：

```
01 public static Integer calc(Integer para) {
02     try {
03         // 模拟一个长时间的执行
04         Thread.sleep(1000);
05     } catch (InterruptedException e) {
06     }
07     return para*para;
08 }
09
10 public static void main(String[] args) throws InterruptedException
11 {
12     CompletableFuture<Void> fu=CompletableFuture.supplyAsync(
```

```
12     .thenApply((i)->Integer.toString(i))
13     .thenApply((str)->"\""+str+"\"")
14     .thenAccept(System.out::println);
15     fu.get();
16 }
```

上述代码中，使用`supplyAsync()`函数执行一个异步任务。接着连续使用流式调用对任务的处理结果进行再加工，直到最后的结果输出。

这里，我们在第15行执行`CompletableFuture.get()`方法，目的是等待`calc()`函数执行完成。如果不进行这个等待调用，由于`CompletableFuture`异步执行的缘故，主函数不等`calc()`方法执行完毕就会退出，随着主线程的结束，所有的`Daemon`线程都会立即退出，从而导致`calc()`方法无法正常完成。

6.5.4 `CompletableFuture`中的异常处理

如果`CompletableFuture`在执行过程中遇到异常，我们可以用函数式编程的风格来优雅地处理这些异常。`CompletableFuture`提供了一个异常处理方法`exceptionally()`：

```
01 public static Integer calc(Integer para) {
02     return para / 0;
03 }
04
05 public static void main(String[] args) throws InterruptedException
```

```
06     CompletableFuture<Void> fu = CompletableFuture
07         .supplyAsync(() -> calc(50))
08         .exceptionally(ex -> {
09             System.out.println(ex.toString());
10             return 0;
11         })
12         .thenApply((i) -> Integer.toString(i))
13         .thenApply((str) -> "\"" + str + "\"")
14         .thenAccept(System.out::println);
15     fu.get();
16 }
```

在上述代码中，第8行对当前的CompletableFuture进行异常处理。如果没有异常发生，则CompletableFuture就会返回原有的结果。如果遇到了异常，就可以在exceptionally()中处理异常，并返回一个默认的值。在上例中，我们忽略了异常堆栈，只是简单地打印异常的信息。

执行上述函数，我们将得到输出：

```
java.util.concurrent.CompletionException: java.lang.ArithmeticExc
"0"
```

6.5.5 组合多个CompletableFuture

CompletableFuture还允许你将多个CompletableFuture进行组合。一种方法是使用thenCompose()，它的签名如下：

```
public <U> CompletableFuture<U> thenCompose(Function<? super T  
CompletionStage<U>> fn)
```

一个CompletableFuture可以在执行完成后，将执行结果通过Function传递给下一个CompletionStage进行处理（Function接口返回新的CompletionStage实例）：

```
01 public static Integer calc(Integer para) {  
02     return para/2;  
03 }  
04  
05 public static void main(String[] args) throws InterruptedException  
06     CompletableFuture<Void> fu =  
07         CompletableFuture.supplyAsync(() -> calc(50))  
08         .thenCompose((i)->CompletableFuture.supplyAsync((  
09         .thenApply((str)->"\" + str + "\"").thenAccept(Sy  
10     fu.get();  
11 }
```

上述代码第8行，将处理后的结果传递给thenCompose()，并进一步传递给后续新生成的CompletableFuture实例。以上代码的输出如下：

```
"12"
```

另外一种组合多个CompletableFuture的方法是thenCombine()，它的签名如下：

```
public <U,V> CompletableFuture<V> thenCombine  
    (CompletionStage<? extends U> other,
```

```
BiFunction<? super T, ? super U, ? extends V> fn)
```

方法thenCombine()首先完成当前CompletableFuture和other的执行。接着，将这两者的执行结果传递给BiFunction（该接口接收两个参数，并有一个返回值），并返回代表BiFunction实例的CompletableFuture对象：

```
01 public static Integer calc(Integer para) {
02     return para / 2;
03 }
04
05 public static void main(String[] args) throws InterruptedException {
06     CompletableFuture<Integer> intFuture = CompletableFuture.supplyAsync(() -> 10);
07     CompletableFuture<Integer> intFuture2 = CompletableFuture.supplyAsync(() -> 10);
08
09     CompletableFuture<Void> fu = intFuture.thenCombine(intFuture2, (i, j) -> {
10         .thenApply((str) -> "\"" + str + "\"")
11         .thenAccept(System.out::println);
12     });
13 }
```

上述代码中，首先生成两个CompletableFuture实例（第6～7行），接着使用thenCombine()组合这两个CompletableFuture，将两者的执行结果进行累加（由第9行的(i, j) -> (i + j)实现），并将其累加结果转为字符串，并输出。上述代码的输出是：

```
"37"
```

6.6 读写锁的改进：StampedLock

StampedLock是Java 8中引入的一种新的锁机制。简单的理解，可以认为它是读写锁的一个改进版本。读写锁虽然分离了读和写的功能，使得读与读之间可以完全并发。但是，读和写之间依然是冲突的。读锁会完全阻塞写锁，它使用的依然是悲观的锁策略，如果有大量的读线程，它也有可能引起写线程的“饥饿”。

而StampedLock则提供了一种乐观的读策略。这种乐观的锁非常类似无锁的操作，使得乐观锁完全不会阻塞写线程。

6.6.1 StampedLock使用示例

StampedLock的使用并不困难，下面是StampedLock的使用示例：

```
01 public class Point {
02     private double x, y;
03     private final StampedLock sl = new StampedLock();
04
05     void move(double deltaX, double deltaY) {           // 这是-
06         long stamp = sl.writeLock();
07         try {
08             x += deltaX;
09             y += deltaY;
10         } finally {
```



```
11         sl.unlockWrite(stamp);
12     }
13 }
14
15 double distanceFromOrigin() {                // 只读方法
16     long stamp = sl.tryOptimisticRead();
17     double currentX = x, currentY = y;
18     if (!sl.validate(stamp)) {
19         stamp = sl.readLock();
20         try {
21             currentX = x;
22             currentY = y;
23         } finally {
24             sl.unlockRead(stamp);
25         }
26     }
27     return Math.sqrt(currentX * currentX + currentY * curr
28 }
29 }
```

上述代码出自JDK的官方文档。它定义了一个点Point类，内部有两个元素x和y，表示点的坐标。第3行，定义了StampedLock锁。第15行定义的distanceFromOrigin()方法是一个只读方法，它只会读取Point的x和y坐标。在读取时，首先使用了StampedLock.tryOptimisticRead()方法。这个方法表示试图尝试一次乐观读。它会返回一个类似于时间戳的邮戳整数stamp。这个stamp就可以作为这一次锁获取的凭证。

接着，在第17行，读取x和y的值。当然，这时我们并不确定这个x和y是否是一致的（在读取x的时候，可能其他线程改写了y的值，使得currentX和currentY处于不一致的状态），因此，我们必须在第18行，使用validate()方法，判断这个stamp是否在读过程发生期间被修改过。如果stamp没有被修改过，则认为这次读取是有效的，因此就可以跳转到第27行，进行数据处理。反之，如果stamp是不可用的，则意味着在读取的过程中，可能被其他线程改写了数据，因此，有可能出现了脏读。如果出现这种情况，我们可以像处理CAS操作那样在一个死循环中一直使用乐观读，直到成功为止。

也可以升级锁的级别。在本例中，我们升级乐观锁的级别，将乐观锁变为悲观锁。在第19行，当判断乐观读失败后，使用readLock()获得悲观的读锁，并进一步读取数据。如果当前对象正在被修改，则读锁的申请可能导致线程挂起。

写入的情况可以参考第5行定义的move()函数。使用writeLock()函数可以申请写锁。这里的含义和读写锁是类似的。

在退出临界区时，不要忘记释放写锁（第11行）或者读锁（第24行）。

可以看到，StampedLock通过引入乐观读来增加系统的并行度。

6.6.2 StampedLock的小陷阱

StampedLock内部实现时，使用类似于CAS操作的死循环反复尝试的策略。在它挂起线程时，使用的是Unsafe.park()函数，而park()函数在

遇到线程中断时，会直接返回（注意，不同于Thread.sleep()，它不会抛出异常）。而在StampedLock的死循环逻辑中，没有处理有关中断的逻辑。因此，这就会导致阻塞在park()上的线程被中断后，会再次进入循环。而当退出条件得不到满足时，就会发生疯狂占用CPU的情况。这一点值得我们注意，下面演示了这个问题：

```
01 public class StampedLockCPUDemo {
02     static Thread[] holdCpuThreads = new Thread[3];
03     static final StampedLock lock = new StampedLock();
04     public static void main(String[] args) throws InterruptedException
05         new Thread() {
06             public void run() {
07                 long readLong = lock.writeLock();
08                 LockSupport.parkNanos(60000000000000L);
09                 lock.unlockWrite(readLong);
10             }
11         }.start();
12     Thread.sleep(100);
13     for (int i = 0; i < 3; ++i) {
14         holdCpuThreads[i] = new Thread(new HoldCPUReadThread());
15         holdCpuThreads[i].start();
16     }
17     Thread.sleep(10000);
18     //线程中断后，会占用CPU
19     for (int i = 0; i < 3; ++i) {
20         holdCpuThreads[i].interrupt();
21     }
```

```

22     }
23
24     private static class HoldCPUReadThread implements Runnable
25     {
26         public void run() {
27             long lockr = lock.readLock();
28             System.out.println(Thread.currentThread().getName() + " acquired read lock");
29             lock.unlockRead(lockr);
30         }
31     }

```

在上述代码中，首先开启线程占用写锁（第7行），注意，为了演示效果，这里使写线程不释放锁而一直等待。接着，开启3个读线程，让它们请求读锁。此时，由于写锁的存在，所有读线程都会被最终挂起。

下面是其中一个读线程在挂起时的信息：

```

"Thread-2" #10 prio=5 os_prio=0 tid=0x14b1d800 nid=0xaafc waiting
  java.lang.Thread.State: WAITING (parking)
    at sun.misc.Unsafe.park(Native Method)
      - parking to wait for  <0x046b54c8> (a java.util.concurrent.locks.StampedLock)
    at java.util.concurrent.locks.StampedLock.acquireRead(StampedLock.java:150)
    at java.util.concurrent.locks.StampedLock.readLock(StampedLock.java:165)
    at geym.conc.ch6.stamped.StampedLockCPUDemo$HoldCPUReadThread.run(StampedLockCPUDemo.java:35)
    at java.lang.Thread.run(Thread.java:745)

```

可以看到，这个线程因为`park()`的操作而进入了等待状态，这种情况是正常的。

而在10秒以后（代码第17行执行了10秒等待），系统中断了这3个读线程，之后，你就会发现，你的CPU占用率极有可能会飙升。这是因为中断导致`park()`函数返回，使线程再次进入运行状态，下面是同一个线程在中断后的信息：

```
"Thread-2" #10 prio=5 os_prio=0 tid=0x14b1d800 nid=0xaafc runnable
  java.lang.Thread.State: RUNNABLE
    at sun.misc.Unsafe.park(Native Method)
      - parking to wait for  <0x046b54c8> (a java.util.concurr
    at java.util.concurrent.locks.StampedLock.acquireRead(Sta
    at java.util.concurrent.locks.StampedLock.readLock(Stampe
    at geym.conc.ch6.stamped.StampedLockCPUDemo$HoldCPUReadTh
  (StampedLockCPUDemo.java:35)
    at java.lang.Thread.run(Thread.java:745)
```

此时，这个线程的状态是**RUNNABLE**，这是我们不愿意看到的。它会一直存在并耗尽CPU资源，直到自己抢占到了锁。

6.6.3 有关**StampedLock**的实现思想

StampedLock的内部实现是基于CLH锁的。CLH锁是一种自旋锁，它保证没有饥饿发生，并且可以保证FIFO（First-In-First-Out）的服务顺序。

CLH锁的基本思想如下：锁维护一个等待线程队列，所有申请锁，但是没有成功的线程都记录在这个队列中。每一个节点（一个节点代表一个线程），保存一个标记位（`locked`），用于判断当前线程是否已经释放锁。

当一个线程试图获得锁时，取得当前等待队列的尾部节点作为其前序节点，并使用类似如下代码判断前序节点是否已经成功释放锁：

```
while (pred.locked) {  
}
```

只要前序节点（`pred`）没有释放锁，则表示当前线程还不能继续执行，因此会自旋等待。

反之，如果前序线程已经释放锁，则当前线程可以继续执行。

释放锁时，也遵循这个逻辑，线程会将自身节点的`locked`位置标记为`false`，那么后续等待的线程就能继续执行了。

如图6.4所示，显示了CLH队列锁的基本思想。

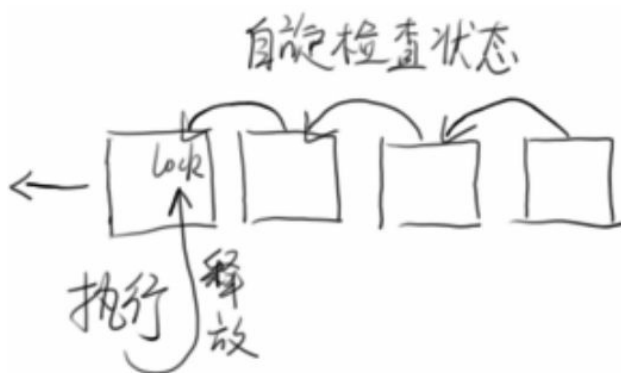


图6-4 CLH队列锁

StampedLock正是基于这种思想，但是实现上更为复杂。

在StampedLock内部，会维护一个等待链表队列：

```
01 /** Wait nodes */
02 static final class WNode {
03     volatile WNode prev;
04     volatile WNode next;
05     volatile WNode cwait;    // 读节点链表
06     volatile Thread thread;  // 当可能被暂停时非空
07     volatile int status;     // 0, WAITING, or CANCELLED
08     final int mode;          // RMODE or WMODE
09     WNode(int m, WNode p) { mode = m; prev = p; }
10 }
11
12 /** CLH 队列头部 */
13 private transient volatile WNode whead;
14 /** CLH 队列尾部 */
15 private transient volatile WNode wtail;
```

上述代码中，WNode为链表的基本元素，每一个WNode表示一个等待线程。字段whead和wtail分别指向等待链表的头部和尾部。

另外一个重要的字段为state：

```
private transient volatile long state;
```

字段state表示当前锁的状态。它是一个long型，有64位，其中，倒数第8位表示写锁状态，如果该位为1，表示当前由写锁占用。

对于一次乐观读的操作，它会执行如下操作：

```
public long tryOptimisticRead() {  
    long s;  
    return (((s = state) & WBIT) == 0L) ? (s & SBITS) : 0L;  
}
```

一次成功的乐观读必须保证当前锁没有写锁占用。其中WBIT用来获取写锁状态位，值为0x80。如果成功，则返回当前state的值（末尾7位清零，末尾7位表示当前正在读取的线程数量）。

如果在乐观读后，有线程申请了写锁，那么state的状态就会改变：

```
1 public long writeLock() {  
2     long s, next; // bypass acquireWrite in fully unlocked cas  
3     return (((s = state) & ABITS) == 0L &&  
4         U.compareAndSwapLong(this, STATE, s, next = s + WB  
5         next : acquireWrite(false, 0L));  
6 }
```

上述代码中第4行，设置写锁位为1（通过加上WBIT（0x80））。这样，就会改变state的取值。那么在乐观锁确认（validate）时，就会发现这个改动，而导致乐观锁失效。

```
public boolean validate(long stamp) {  
    U.loadFence();  
    return (stamp & SBITS) == (state & SBITS);  
}
```

上述validate()函数比较当前stamp和发生乐观锁时取得的stamp，如果不一致，则宣告乐观锁失败。

乐观锁失败后，则可以提升锁级别，使用悲观读锁。

```
1 public long readLock() {
2     long s = state, next; // bypass acquireRead on common unco
3     return ((whead == wtail && (s & ABITS) < RFULL &&
4             U.compareAndSwapLong(this, STATE, s, next = s + RU
5             next : acquireRead(false, 0L)));
6 }
```

悲观读会尝试设置state状态（第4行），它会将state加1（前提是读线程数量没有溢出，对于读线程数量溢出的情况，会使用辅助的readerOverflow进行统计，我们在这里不做过于烦琐的讨论），用于统计读线程的数量。如果失败，则进入acquireRead()二次尝试锁获取。

在acquireRead()中，线程会在不同条件下进行若干次自旋，试图通过CAS操作获得锁。如果自旋宣告失败，则会启用CLH队列，将自己加到队列中。之后再进行自旋，如果发现自己成功获得了读锁，则会进一步把自己cwait队列中的读线程全部激活（使用Unsafe.unpark()方法）。如果最终依然无法成功获得读锁，则会使用Unsafe.park()方法挂起当前线程。

方法acquireWrite()和acquireRead()也非常类似，也是通过自旋尝试、加入等待队列、直至最终Unsafe.park()挂起线程的逻辑进行的。释放锁时与加锁动作相反，以unlockWrite()为例：

```
1 public void unlockWrite(long stamp) {
2     WNode h;
3     if (state != stamp || (stamp & WBIT) == 0L)
```

```
4         throw new IllegalMonitorStateException();
5     state = (stamp += WBIT) == 0L ? ORIGIN : stamp;
6     if ((h = whead) != null && h.status != 0)
7         release(h);
8 }
```

上述代码第5行，将写标记位清零，如果state发生溢出，则退回到初始值。

接着，如果等待队列不为空，则从等待队列中激活一个线程（绝大部分情况下是第1个等待线程）继续执行（第7行）。

6.7 原子类的增强

在之前的章节中已经提到了原子类的使用，无锁的原子类操作使用系统的CAS指令，有着远远超越锁的性能。那是否有可能在性能上更上一层楼呢？答案是肯定的。在Java 8中引入了LongAdder类，这个类也在java.util.concurrent.atomic包下，因此，可以推测，它也是使用了CAS指令。

6.7.1 更快的原子类：LongAdder

大家对AtomicInteger的基本实现机制应该比较了解。它们都是在一个死循环内，不断尝试修改目标值，直到修改成功。如果竞争不激烈，那么修改成功的概率就很高，否则，修改失败的概率就很高。在大量修改失败时，这些原子操作就会进行多次循环尝试，因此性能就会受到影响。

那么当竞争激烈的时候，我们应该如何进一步提高系统的性能呢？一种基本方案就是可以使用热点分离，将竞争的数据进行分解，基于这个思路，大家应该可以想到一种对传统AtomicInteger等原子类的改进方法。虽然在CAS操作中没有锁，但是像减小锁粒度这种分离热点的思想依然可以使用。一种可行的方案就是仿造ConcurrentHashMap，将热点数据分离。比如，可以将AtomicInteger的内部核心数据value分离成一个数组，每个线程访问时，通过哈希等算法映射到其中一个数字进行计数，而最终的计数结果，则为这个数组的求和累加，如图6.5所示，显示了这种优化思路。其中，热点数据value被分离成多个单元cell，每个

cell独自维护内部的值，当前对象的实际值由所有的cell累计合成，这样，热点就进行了有效的分离，提高了并行度。LongAdder正是使用了这种思想。

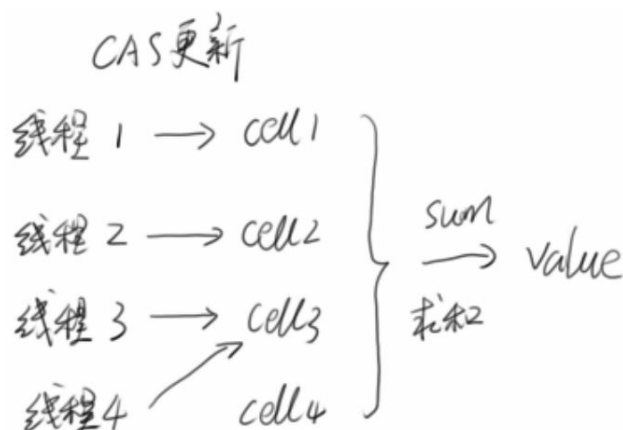


图6.5 原子类的优化思路

在实际的操作中，LongAdder并不会一开始就动用数组进行处理，而是将所有数据都先记录在一个称为base的变量中。如果在多线程条件下，大家修改base都没有冲突，那么也没有必要扩展为cell数组。但是，一旦base修改发生冲突，就会初始化cell数组，使用新的策略。如果使用cell数组更新后，发现在某一个cell上的更新依然发生冲突，那么系统就会尝试创建新的cell，或者将cell的数量加倍，以减少冲突的可能。

下面我们简单分析一下increment()方法（该方法会将LongAdder自增1）的内部实现：

```
01 public void increment() {
02     add(1L);
03 }
04 public void add(long x) {
```

```

05     Cell[] as; long b, v; int m; Cell a;
06     if ((as = cells) != null || !casBase(b = base, b + x)) {
07         boolean uncontended = true;
08         if (as == null || (m = as.length - 1) < 0 ||
09             (a = as[getProbe() & m]) == null ||
10             !(uncontended = a.cas(v = a.value, v + x)))
11             longAccumulate(x, null, uncontended);
12     }
13 }

```

它的核心是第4行的add()方法。最开始cells为null，因此数据会向base增加（第6行）。但是如果对base的操作冲突，则会进入第7行，并设置冲突标记uncontended为true。接着，如果判断cells数组不可用，或者当前线程对应的cell为null，则直接进入longAccumulate()方法。否则会尝试使用CAS方法更新对应的cell数据，如果成功，则退出，失败则进入longAccumulate()方法。

由于longAccumulate()方法比较复杂，限于篇幅，这里不再展开讨论，其大致内容是根据需要创建新的cell或者对cell数组进行扩容，以减少冲突。

下面，让我们简单地对LongAddr、原子类以及同步锁进行性能测试。测试方法是使用多个线程对同一个整数进行累加，观察使用3种不同方法时所消耗的时间。

首先，我们定义一些辅助变量：

```

private static final int MAX_THREADS = 3; //线程数

```

```

private static final int TASK_COUNT = 3;                //任务数
private static final int TARGET_COUNT = 100000000;      //目标总数

private AtomicLong account =new AtomicLong(0L);        //无锁的原子
private LongAdder lacount=new LongAdder();
private long count=0;

static CountdownLatch cdlsync=new CountdownLatch(TASK_COUNT);
static CountdownLatch cdlatomic=new CountdownLatch(TASK_COUNT);
static CountdownLatch cdladdr=new CountdownLatch(TASK_COUNT);

```

上述代码中，指定了测试线程数量、目标总数以及3个初始值为0的整型变量account、lacount和count。它们分别表示使用AtomicLong、LongAdder和锁进行同步时的操作对象。

下面是使用同步锁时的测试代码：

```

01 protected synchronized long inc(){                  //有锁的
02     return ++count;
03 }
04
05 protected synchronized long getCount(){             //有锁的
06     return count;
07 }
08
09
10 public class SyncThread implements Runnable{
11     protected String name;

```

```
12     protected long starttime;
13     LongAdderDemo out;
14     public SyncThread(LongAdderDemo o,long starttime){
15         out=o;
16         this.starttime=starttime;
17     }
18     @Override
19     public void run() {
20         long v=out.getCount();
21         while(v<TARGET_COUNT){                //在到达
22             v=out.inc();
23         }
24         long endtime=System.currentTimeMillis();
25         System.out.println("SyncThread spend:"+(endtime-starttime));
26         cdsync.countDown();
27     }
28 }
29
30 public void testSync() throws InterruptedException{
31     ExecutorService exe=Executors.newFixedThreadPool(MAX_THREADS);
32     long starttime=System.currentTimeMillis();
33     SyncThread sync=new SyncThread(this,starttime);
34     for(int i=0;i<TASK_COUNT;i++){
35         exe.submit(sync);                    //提交线程
36     }
37     cdsync.await();
38     exe.shutdown();
```

39 }

上述代码第10行，定义线程SyncThread，它使用加锁方式增加count的值。在第30行定义的testSync()方法中，使用线程池控制多线程进行累加操作。

使用类似的方法实现原子类累加计时统计：

```
01 public class AtomicThread implements Runnable{
02     protected String name;
03     protected long starttime;
04     public AtomicThread(long starttime){
05         this.starttime=starttime;
06     }
07     @Override
08     public void run() { //在3
09         long v=acount.get();
10         while(v<TARGET_COUNT){
11             v=acount.incrementAndGet(); //无4
12         }
13         long endtime=System.currentTimeMillis();
14         System.out.println("AtomicThread spend:"+(endtime-star
15         cdlatomic.countDown());
16     }
17 }
18
19 public void testAtomic() throws InterruptedException{
20     ExecutorService exe=Executors.newFixedThreadPool(MAX_THREA
```



```

21     long starttime=System.currentTimeMillis();
22     AtomicThread atomic=new AtomicThread(starttime);
23     for(int i=0;i<TASK_COUNT;i++){
24         exe.submit(atomic);                //提
25     }
26     cdlatomic.await();
27     exe.shutdown();
28 }

```

同理，以下代码使用LongAddr实现类似的功能：

```

01 public class LongAddrThread implements Runnable{
02     protected String name;
03     protected long starttime;
04     public LongAddrThread(long starttime){
05         this.starttime=starttime;
06     }
07     @Override
08     public void run() {
09         long v=lacount.sum();
10         while(v<TARGET_COUNT){
11             lacount.increment();
12             v=lacount.sum();
13         }
14         long endtime=System.currentTimeMillis();
15         System.out.println("LongAdder spend:"+(endtime-startti
16         cdladdr.countDown());

```

```

17     }
18 }
19
20 public void testAtomicLong() throws InterruptedException{
21     ExecutorService exe=Executors.newFixedThreadPool(MAX_THREA
22     long starttime=System.currentTimeMillis();
23     LongAddrThread atomic=new LongAddrThread(starttime);
24     for(int i=0;i<TASK_COUNT;i++){
25         exe.submit(atomic);                //提
26     }
27     cdladdr.await();
28     exe.shutdown();
29 }

```

注意，由于LongAddr中，将单个数值分解为多个不同的段。因此，在进行累加后，上述代码中第11行的increment()函数并不能返回当前的数值。要取得当前的实际值，需要使用第12行的sum()函数重新计算。这个计算是需要有额外的成本的，但即使加上这个额外成本，LongAddr的表现还是比AtomicLong要好。

执行这些代码，就可以得到锁、原子类和LongAddr三者的性能比较数据，如下所示：

```

SyncThread spend:1784ms v=10000002
SyncThread spend:1784ms v=10000000
SyncThread spend:1784ms v=10000001
AtomicThread spend:695ms v=10000001
AtomicThread spend:695ms v=10000000

```

```
AtomicThread spend:695ms v=10000002
LongAdder spend:227ms v=10000002
LongAdder spend:227ms v=10000002
LongAdder spend:227ms v=10000002
```

可以看到，就计数性能而言，LongAdder已经超越了普通的原子操作了。其中，锁操作耗时约1784ms，普通原子操作耗时约695ms，而LongAdder仅需要227ms左右。

LongAdder的另外一个优化手段是避免了伪共享。大家可以先回顾一下第5章中有关伪共享的问题。但是，需要注意的是，LongAdder中并不是直接使用padding这种看起来比较碍眼的做法，而是引入了一种新的注释“@sun.misc.Contended”。

对于LongAdder中的每一个Cell，它的定义如下所示：

```
@sun.misc.Contended
static final class Cell {
    volatile long value;
    Cell(long x) { value = x; }
    final boolean cas(long cmp, long val) {
        return UNSAFE.compareAndSwapLong(this, valueOffset, cmp,
    }
    省略其他不必要的信息
```

可以看到，在上述代码第1行声明了Cell类为sun.misc.Contended。这将会使得Java虚拟机自动为Cell解决伪共享问题。

当然，在我们自己的代码中也可以使用sun.misc.Contended来解决伪

共享问题，但是需要额外使用虚拟机参数-XX:-RestrictContended，否则，这个注释将被忽略。

大家应该还记得第5章中有关伪共享的案例吧！限于篇幅，这里不再贴出完整代码，只给出关键部分的改动。我们将VolatileLong修改如下：

```
@sun.misc.Contended
public final static class VolatileLong {
    public volatile long value = 0L;
}
```

在这里，我们去除了那些看起来不太雅观的padding，同时增加了sun.misc.Contended申明，这个就告诉虚拟机我们希望在这个类上解决伪共享问题。然后，我们就可以测试这段代码了。当然了，千万不要忘记指定虚拟机参数-XX:-RestrictContended，否则，你的这个优化将被无视。

跑一下优化后的程序，是不是比传统的方式快很多呢？

6.7.2 LongAdder的功能增强版：LongAccumulator

LongAccumulator是LongAdder的亲兄弟，它们有公共的父类Striped64。因此，LongAccumulator内部的优化方式和LongAdder是一样的。它们都将一个long型整数进行分割，存储在不同的变量中，以防止多线程竞争。两者的主要逻辑是类似的，但是LongAccumulator是

LongAdder的功能扩展，对于LongAdder来说，它只是每次对给定的整数执行一次加法，而LongAccumulator则可以实现任意函数操作。

可以使用下面的构造函数创建一个LongAccumulator实例：

```
public LongAccumulator(LongBinaryOperator accumulatorFunction, long
```

第1个参数accumulatorFunction就是需要执行的二元函数（接收两个long形参数并返回long），第2个参数是初始值。

下面这个例子展示了LongAccumulator的使用，它将通过多线程访问若干个整数，并返回遇到的最大的那个数字。

```
01 public static void main(String[] args) throws Exception {
02     LongAccumulator accumulator = new LongAccumulator(Long::ma
03     Thread[] ts = new Thread[1000];
04
05     for (int i = 0; i < 1000; i++) {
06         ts[i] = new Thread(() -> {
07             Random random = new Random();
08             long value = random.nextLong();
09             accumulator.accumulate(value);
10         });
11         ts[i].start();
12     }
13     for (int i = 0; i < 1000; i++) {
14         ts[i].join();
15     }
```

```
16      System.out.println(accumulator.longValue());  
17 }
```

上述代码第2行，构造了LongAccumulator实例。因为我们要过滤最大值，因此传入Long::max函数句柄。当有数据通过accumulate()方法传入LongAccumulator后（第9行），LongAccumulator会通过Long::max识别最大值并且保存在内部（很可能是cell数组内，也可能是base）。在代码第16行，通过longValue()函数对所有的cell进行Long::max操作，得到最大值。

6.8 参考文献

- 对于函数式编程的字节码分析部分
 - 《实战Java虚拟机：JVM故障诊断与性能优化》
- 有关CompletableFuture的使用
 - <http://www.javacodegeeks.com/2013/05/java-8-definitive-guide-to-completablefuture.html>
- StampedLock的官方文档
 - <https://docs.oracle.com/javase/8/docs/api/java/util/concurrent/locks/>
- StampedLock使用不慎导致CPU占用率偏高的解决方案
 - <http://feed.hjue.me/articles/detail/2014-07-10/364485/cpu-bug-concurrency>
- CLH自旋锁和其他自旋锁的详细介绍
 - 《The Art of Multiprocessor Programming》
- JDK 8中的伪共享问题
 - <http://www.programering.com/a/MDM5IzNwATg.html>

第7章 使用Akka构建高并发程序

我们知道，写出一个正确的、高性能并且可扩展的并发程序是相当困难的，那么是否有一个好的框架可以帮助我们轻松构建这么一个应用呢？答案是肯定的，那就是Akka。Akka是一款遵循Apache 2许可的开源人员，这意味着你可以无偿并且几乎没有限制地使用它，包括将它应用于商业环境中。

Akka是用Scala创建的，但由于Scala和Java一样，都是Java虚拟机上的语言，本质上说，两者并没有什么不同，因此，我们也可以在Java中使用Akka。考虑到Java开发人员的数量远远高于Scala，为了方便大众，在这里，我将全程使用Java来作为Akka的宿主语言（本书使用Akka 2.11-2.3.7作为演示）。但我并不打算在这里把对Akka的介绍写成一个Akka使用手册，因此，不会对Akka进行全方位完整的API介绍。只是希望在这里对Akka的主要功能进行简单的描述，帮助大家尽快理解Akka的基本思想。

那么使用Akka能够给我们带来什么好处呢？

首先Akka提供了一种称为Actor的并发模型，其粒度比线程更小，这意味着你可以在系统中启用极其大量的Actor。

其次，Akka中提供了一套容错机制，允许在Actor出现异常时进行一些恢复或者重置操作。

最后，通过Akka不仅可以在单机上构建高并发程序，也可以在网络中构建分布式程序，并提供位置透明的Actor定位服务。

下面就让我们正式开启Akka之旅吧！

7.1 新并发模型：Actor

对于并发程序来说，线程始终作为并发程序的基本执行单元。但在Akka中，你可以完全忘记线程了。当你使用Akka时，你就有一个全新的执行单元——Actor。Actor是什么呢？

简单来说，你可以把Actor比喻成一个人。多个人之间可以使用语言进行交流。比如，老师问同学5乘以5是多少呀？同学听到问题后，想了想，回答说是25。Actor之间的通信方式和上述对话形式几乎是一模一样的。

传统Java并程序，还是完全基于面向对象的方法。我们还是通过对象的方法调用进行信息的传递。这时，如果对象的方法会修改对象本身的状态，那么在多线程情况下，就有可能出现对象状态的不一致，所以我们必须对这类方法调用进行同步。当然，同步往往就是以牺牲性能为代价的。

在Actor模型中，我们失去了对象的方法调用，我们并不是通过调用Actor对象的某一个方法来告诉Actor你需要做什么，而是给Actor发送一条消息。当一个Actor收到消息后，它有可能会根据消息的内容做出某些行为，包括更改自身状态。但是，在这种情况下，这个状态的更改是Actor自己进行的，并不是由外界被迫进行的。

7.2 Akka之Hello World

在了解了Actor的基本行为模式后，我们通过简单的Hello World程序来进一步了解一下Akka的开发。

首先让我们看一下，第1个Actor的实现：

```
01 public class Greeter extends UntypedActor {
02     public static enum Msg {
03         GREET, DONE;
04     }
05
06     @Override
07     public void onReceive(Object msg) {
08         if (msg == Msg.GREET) {
09             System.out.println("Hello World!");
10             getSender().tell(Msg.DONE, getSelf());
11         } else
12             unhandled(msg);
13     }
14 }
```

上述代码中，定义了一个欢迎者（Greeter）Actor，它继承自UntypedActor（它自然就是Akka中的核心成员了）。UntypedActor就是我们所说的Actor，之所以这里强调是无类型的，那是因为在Akka中，还支持一种有类型的Actor。有类型的Actor可以使用系统中的其他类型

构造，可以缓解Java单继承的问题。因为你在继承了UntypedActor后，就不能再继承系统中的其他类了。如果你一定想这么做，那么就只能选择有类型的Actor。否则，UntypedActor应该就是你的首选。

在这里，代码第2~4行，定义了消息类型。这里只有两种类型，欢迎（GREET）以及完成（DONE）。当Greeter收到GREET消息时，就会在控制台打印“Hello World”，并且向消息发送方发送DONE信息（第10行）。

与Greeter交流的另外一个Actor是HelloWorld，它的实现如下：

```
01 public class HelloWorld extends UntypedActor {
02     ActorRef greeter;
03
04     @Override
05     public void preStart() {
06         greeter = getContext().actorOf(Props.create(Greeter.cl
07         System.out.println("Greeter Actor Path:" + greeter.pat
08         greeter.tell(Greeter.Msg.GREET, getSelf());
09     }
10
11     @Override
12     public void onReceive(Object msg) {
13         if (msg == Greeter.Msg.DONE) {
14             greeter.tell(Greeter.Msg.GREET, getSelf());
15             getContext().stop(getSelf());
16         } else
17             unhandled(msg);
```

```
18     }  
19 }
```

上述代码实现了一个名为HelloWorld的Actor。第5行的preStart()方法为Akka的回调方法，在Actor启动前，会被Akka框架调用，完成一些初始化的工作。在这里，我们在HelloWorld中创建了Greeter的实例（第6行），并且向它发送GREET消息（第8行）。此时，由于创建Greeter时使用的是HelloWorld的上下文，因此，它属于HelloWorld的子Actor。

第12行定义的onReceive()函数为HelloWorld的消息处理函数。在这里，只处理DONE的消息。在收到DONE消息后，它会再向Greeter发送GREET消息，接着将自己停止。

因此，Greeter会前后收到两条GREET消息，会打印两次“Hello World”。

最后，让我们看一下主函数main()：

```
1 public class HelloMainSimple {  
2     public static void main(String[] args) {  
3         ActorSystem system = ActorSystem.create("Hello", ConfigFacto  
4             ActorRef a = system.actorOf(Props.create(HelloWorld.cla  
5             System.out.println("HelloWorld Actor Path:" + a.path()  
6     }  
7 }
```

程序第3行，创建了ActorSystem，表示管理和维护Actor的系统。一般来说，一个应用程序只需要一个ActorSystem就够用了。

ActorSystem.create()的第1个参数“Hello”为系统名称，第2个参数为配置

文件。

第4行通过ActorSystem创建一个顶级的Actor（HelloWorld）。

配置文件samplehello.conf的内容如下：

```
akka {  
    loglevel = INFO  
}
```

在这里，只是简单地配置了一下日志级别为INFO。

执行上述代码，可以看到以下输出：

```
1 HelloWorld Actor Path:akka://Hello/user/helloWorld  
2 Greeter Actor Path:akka://Hello/user/helloWorld/greeter  
3 Hello World!  
4 Hello World!  
5 [INFO] [05/13/2015 21:15:01.299] [Hello-akka.actor.default-disp  
[akka://Hello/user/helloWorld] Message [geym.akka.demo.hello.Gree  
Actor[akka://Hello/user/helloWorld/greeter#-1698722495] to  
Actor[akka://Hello/user/helloWorld#-1915075849] was not delivered  
encountered. This logging can be turned off or adjusted with conf  
'akka.log-dead-letters' and 'akka.log-dead-letters-during-shutdown'
```

第1行打印了HelloWorld Actor的路径。它是系统内第1个被创建的Actor。它的路径为：akka://Hello/user/helloWorld。其中第1个Hello表示ActorSystem的系统名，可以看一下我们构造这ActorSystem时，传入的第1个参数就是Hello。接着user表示用户Actor。所有的用户Actor都会挂

载在user这个路径下。第3个helloWorld就是这个Actor的名字。

同理，第2个Greeter Actor的路径结构和HelloWorld是完全一致的。输出的第3、4行显示了Greeter打印的两条信息。第5行表示系统遇到了一条消息投递失败，失败的原因是HelloWorld将自己终止了，导致Greeter发送的信息无法投递。

可以看到，当使用Actor进行并行程序开发时，我们的关注点已经不在线程上了。实际上，线程调度已经被Akka框架进行封装，我们只需要关注Actor对象即可。而Actor对象之间的交流和普通的对象的函数调用有明显区别。它们是通过显示的消息发送来传递信息的。

当系统内有多个Actor存在时，Akka会自动在线程池中选择线程来执行我们的Actor。因此，多个不同的Actor有可能会被同一个线程执行，同时，一个Actor也有可能被不同的线程执行。因此，一个值得注意的地方是：不要在一个Actor中执行耗时的代码，这样可能会导致其他Actor的调度出现问题。

7.3 有关消息投递的一些说明

整个Akka应用是由消息驱动的。消息是除了Actor之外最重要的核心组件。作为在并发程序中的核心组件，在Actor之间传递的消息应该满足不可变性，也就是不变模式。因为可变的对象无法高效的在并发环境中使用。理论上Akka中的消息可以使用任何对象实例，但实际使用中，强烈推荐使用不可变的对象。一个典型的不可变对象的实现如下：

```
01 public final class ImmutableMessage {
02     private final int sequenceNumber;
03
04     private final List<String> values;
05
06     public ImmutableMessage(int sequenceNumber, List<String>
07         this.sequenceNumber = sequenceNumber;
08         this.values = Collections.unmodifiableList(new ArrayLi
09     }
10
11     public int getSequenceNumber() {
12         return sequenceNumber;
13     }
14
15     public List<String> getValues() {
16         return values;
17     }
```


上述代码实现了一个不可变的消息。注意代码中对`final`的使用，它声明了当前消息中的几个字段都是常量，在消息构造完成后，就不能再发生改变了。更加需要注意的是，对于`values`字段，`final`关键字只能保证`values`引用的不可变性，并无法保证`values`对象的不可变性。为了实现彻底的不可变性，代码第8行构造了一个不可变的`List`对象。

对于消息投递，大家可能还有另外一个疑问，那就是消息投递究竟是以何种策略进行的呢？也就是发出去的消息一定会被对方接收到吗？如果接收不到会重发吗？有没有可能重复接收消息呢？

实际上，对于消息投递，我们可以有3种不同的策略：

第1种，称为至多一次投递。在这种策略中，每一条消息最多会被投递一次。在这种情况下，可能偶尔会出现消息投递失败，而导致消息丢失。

第2种称为至少一次投递。在这种策略中，每一条消息至少会被投递一次，直到成功为止。因此在一些偶然的场合，接受者可能会收到重复的消息，但不会发生消息丢失。

第3种称为精确的消息投递。也就是所有的消息保证被精确地投递并成功接收一次，既不会有丢失，也不会有重复接收。

很明显，第1种策略是最高性能，最低成本的。因为系统只要负责把消息送出去就可以了，不需要关注是否成功。第2种策略则需要保存消息投递的状态并不断充实。而第3种策略则是成本最高且最不容易实现的。

那我们是否真的需要保证消息投递的可靠性呢？

答案是否定的。实际上，我们没有必要在Akka层保证消息的可靠性。这样做，成本太高了，也是没有必要的。消息的可靠性更应该在应用的业务层去维护，因为也许在有些时候，丢失一些消息完全是符合应用要求的。因此，在使用Akka时，需要在业务层对此进行保证。

此外，对于消息投递Akka可以在一定程度上保证顺序性。比如，Actor A1向A2顺序发送了M1、M2和M3三条消息。Actor A3向A2顺序发送了M4、M5和M6三条消息。那么系统可以保证：

- (1) 如果M1没有丢失，那它一定先于M2和M3被A2收到。
- (2) 如果M2没有丢失，那它一定先于M3被A2收到。
- (3) 如果M4没有丢失，那它一定先于M5和M6被A2收到。
- (4) 如果M5没有丢失，那它一定先于M6被A2收到。
- (5) 对A2来说，来自A1和A3的消息可能交织在一起，没有顺序保证。

在这里，值得注意的一点是，这种消息投递规则不具备可传递性，比如：

Actor A向C发送了M1，接着，Actor A向B发送了M2，B将M2转发给Actor C。那么在这种情况下，C收到M1和M2的先后顺序是没有保证的。

7.4 Actor的生命周期

Actor在系统中产生后，也存在着“生老病死”的活动周期。Akka框架提供了若干回调函数，让我们得以在Actor的活动周期内进行一些业务相关的行为。Actor的生命周期如图7.1所示。

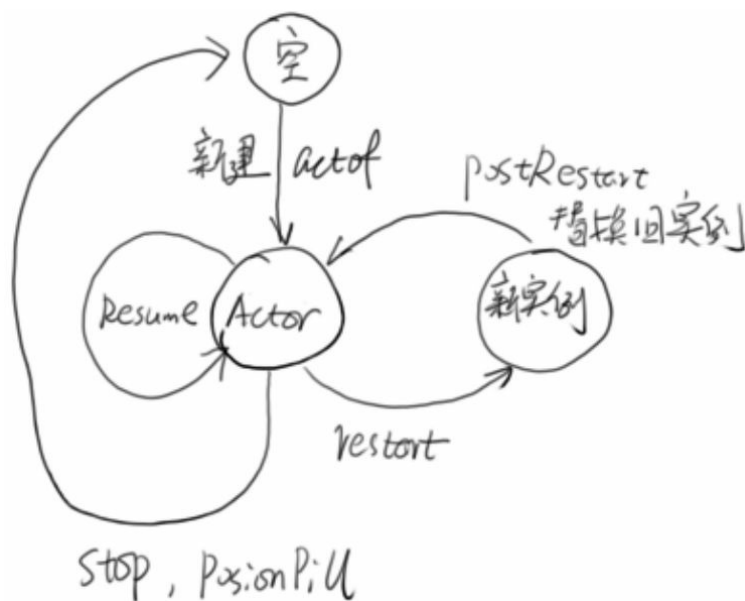


图7.1 Actor的生命周期

一个Actor在actorOf()函数被调用后开始建立，Actor实例创建后，会回调preStart()方法。在这个方法里，我们可以进行一些资源的初始化工作。在Actor的工作过程中，可能会出现一些异常，这种情况下，Actor会需要重启。当Actor被重启时，会回调preRestart()方法（在老的实例上），接着系统会创建一个新的Actor对象实例（虽然是新的实例，但它们都表示同一个Actor）。当新的Actor实例创建后，会回调postRestart()方法，表示启动完成，同时新的实例将会代替旧的实例。停止一个Actor也有很多方式，你可以调用stop()方法或者给Actor发送一个

PosionPill（毒药丸）。Actor停止时，postStop()方法会被调用，同时这个Actor的监视者会收到一个Terminated消息。

下面让我们建立一个带有生命周期回调函数的Actor：

```
public class MyWorker extends UntypedActor {
    private final LoggingAdapter log = Logging.getLogger(getContext());
    public static enum Msg {
        WORKING, DONE, CLOSE;
    }
    @Override
    public void preStart(){
        System.out.println("MyWorker is starting");
    }
    @Override
    public void postStop(){
        System.out.println("MyWorker is stopping");
    }
    @Override
    public void onReceive(Object msg) {
        if (msg == Msg.WORKING) {
            System.out.println("I am working");
        }
        if (msg == Msg.DONE) {
            System.out.println("Stop working");
        }
        if (msg == Msg.CLOSE) {
            System.out.println("I will shutdown");
        }
    }
}
```

```

        getSender().tell(Msg.CLOSE, getSelf());
        getContext().stop(getSelf());
    } else
        unhandled(msg);
}
}

```

上述代码定义了一个名为MyWorker的Actor。它重载了preStart()和postStop()两个方法。一般来说，我们可以使用preStart()来初始化一些资源，使用postStop()来进行资源的释放。这个Actor很简单，当它收到WORKING消息时，就打印“I am working”，收到DONE消息时，打印“Stop working”。

接着，我们为MyWorker指定一个监视者，监视者就如同一个劳动监工，一旦MyWorker因为意外停止工作，监视者就会收到一个通知。

```

01 public class WatchActor extends UntypedActor {
02     private final LoggingAdapter log = Logging.getLogger(getCo
03
04     public WatchActor(ActorRef ref) {
05         getContext().watch(ref);
06     }
07
08     @Override
09     public void onReceive(Object msg) {
10         if (msg instanceof Terminated) {
11             System.out.println(String.format("%s has terminate
12                 ((Terminated) msg).getActor().path()));

```

```

13         getContext().system().shutdown();
14     } else {
15         unhandled(msg);
16     }
17 }
18 }

```

上述代码定义了一个监视者WatchActor，它本质上也是一个Actor，但不同的是，它会在它的上下文中watch一个Actor（第5行）。如果将来这个被监视的Actor的退出终止，WatchActor就能收到一条Terminated消息（代码第10行）。在这里，我们将简单地打印终止消息Terminated中的相关Actor路径，并且关闭整个ActorSystem（第13行）。

主函数如下：

```

01 public class DeadMain {
02     public static void main(String[] args) {
03         ActorSystem system = ActorSystem
04             .create("deadwatch", ConfigFactory.load("sample.conf"));
05         ActorRef worker = system.actorOf(Props.create(MyWorker.class));
06         system.actorOf(Props.create(WatchActor.class, worker),
07             "watcher");
08         worker.tell(MyWorker.Msg.WORKING, ActorRef.noSender());
09         worker.tell(MyWorker.Msg.DONE, ActorRef.noSender());
10         worker.tell(PoisonPill.getInstance(), ActorRef.noSender());
11     }
12 }

```

上述代码中，我们首先创建ActorSystem全局实例（第3~4行），

接着创建MyWorker Actor和WatchActor。注意第6行的Props.create()方法，它的第1个参数为要创建的Actor类型，第2个参数为这个Actor的构造函数的参数（在这里，就是要调用WatchActor的构造函数）。接着，向MyWorker先后发送WORKING和DONE两条消息。最后在第9行，发送一条特殊的消息PoisonPill。PoisonPill就是毒药丸，它会直接毒死接收方，让其终止。

执行上述代码，系统输出如下：

```
MyWorker is starting
I am working
Stop working
MyWorker is stopping
akka:///deadwatch/user/worker has terminated, shutting down system
```

从这个输出中可以看到，MyWorker生命周期中的两个回调函数以及消息处理函数都被正常调用。最后一行输出也显示WatchActor正常监视到MyWorker的终止。

7.5 监督策略

如果一个Actor在执行过程中发生意外，比如没有处理某些异常，导致出错，那么这个时候应该怎么办呢？系统是应该当做什么都没发生过，继续执行，还是认为遇到了一个系统性的错误而重启Actor甚至是它所有的兄弟Actor呢？

对于这种情况，Akka框架给予了我们足够的控制权。在Akka框架内，父Actor可以对子Actor进行监督，监控Actor的行为是否有异常。大体上，监督策略可以分为两种：一种是OneForOneStrategy的监督，另外一种是全ForOneStrategy。

对于OneForOneStrategy的策略，父Actor只会对出问题的子Actor进行处理，比如重启或者停止，而对于AllForOneStrategy，父Actor会对出问题的子Actor以及它所有的兄弟都进行处理。很显然，对于AllForOneStrategy策略，它更加适合于各个Actor联系非常紧密的场景，如果多个Actor间只要有一个Actor出现故障，则宣告整个任务失败，就比较适合使用AllForOneStrategy，否则，在更多的场景中，应该使用OneForOneStrategy。当然了，OneForOneStrategy也是Akka的默认策略。

在一个指定的策略中，我们可以对Actor的失败情况进行相应的处理，比如：当失败时，我们可以无视这个错误，继续执行Actor，就像什么事都没发生过一样。或者可以重启这个Actor，甚至可以让这个Actor彻底停止工作。要指定这些监督行为，只要构造一个自定义的监督策略即可。

下面让我们简单看一下SupervisorStrategy的使用和设置。首先，需要定一个父Actor，它作为所有子Actor的监督者：

```
01 public class Supervisor extends UntypedActor {
02     private static SupervisorStrategy strategy = new OneForOneStra
03         new Function<Throwable, Directive>() {
04             @Override
05             public Directive apply(Throwable t) {
06                 if (t instanceof ArithmeticException) {
07                     System.out.println("meet ArithmeticExc
08                     return SupervisorStrategy.resume();
09                 } else if (t instanceof NullPointerException)
10                     System.out.println("meet NullPointerException
11                     return SupervisorStrategy.restart();
12                 } else if (t instanceof IllegalArgumentExceptionExc
13                     return SupervisorStrategy.stop();
14                 } else {
15                     return SupervisorStrategy.escalate();
16                 }
17             }
18         });
19
20     @Override
21     public SupervisorStrategy supervisorStrategy() {
22         return strategy;
23     }
24 }
```

```
25     public void onReceive(Object o) {
26         if (o instanceof Props) {
27             getContext().actorOf((Props) o, "restartActor");
28         } else {
29             unhandled(o);
30         }
31     }
32 }
```

上述代码第2~18行，定义了一个OneForOneStrategy的监督策略。在这个监督策略中，运行Actor在遇到错误后，在1分钟内进行3次重试。如果超过这个频率，那么就会直接杀死Actor。具体的策略由第5~16行定义。这里的含义是，当遇到ArithmeticException异常时（比如除以0的错误），继续指定这个Actor，不做任何处理（第8行）；当遇到空指针时，进行Actor的重启（第11行）。如果遇到IllegalArgumentException异常，则直接停止Actor（第13行）。对于在这个函数中没有涉及的异常，则向上抛出，由更顶层的Actor处理（第15行）。

第20~23行覆盖父类的supervisorStrategy()方法，设置使用自定义的监督策略。

第27行用来新建一个名为restartActor的子Actor，这个子Actor就由当前的Supervisor进行监督了。当Supervisor接收一个Props对象时，就会根据这个Props配置生成一个restartActor。

RestartActor的实现如下：

```
01 public class RestartActor extends UntypedActor {
02     public enum Msg {
03         DONE, RESTART
04     }
05
06     @Override
07     public void preStart() {
08         System.out.println("preStart hashCode:" + this.hashCode()
09     }
10
11     @Override
12     public void postStop() {
13         System.out.println("postStop hashCode:" + this.hashCode()
14     }
15
16     @Override
17     public void postRestart(Throwable reason) throws Exception {
18         super.postRestart(reason);
19         System.out.println("postRestart hashCode:" + this.hashCode()
20     }
21
22     @Override
23     public void preRestart(Throwable reason, Option opt) throws Exception {
24         System.out.println("preRestart hashCode:" + this.hashCode()
25     }
26
27     @Override
```

```

28     public void onReceive(Object msg) {
29         if (msg == Msg.DONE) {
30             getContext().stop(getSelf());
31         } else if (msg == Msg.RESTART) {
32             System.out.println(((Object)null).toString());
33             //抛出异常 默认会被restart，但这里会resume
34             double a = 0 / 0;
35         }
36         unhandled(msg);
37     }
38 }

```

第6~25行，定义了一些Actor的生命周期的回调接口。目的是更好地观察Actor的活动情况。在第32~34行模拟了一些异常情况，第32行会抛出NullPointerException，而第34行因为除以零，所以会抛出ArithmeticException。

主函数如下定义：

```

01 public static void customStrategy(ActorSystem system){
02     ActorRef a = system.actorOf(Props.create(Supervisor.class)
03     a.tell(Props.create(RestartActor.class), ActorRef.noSender
04
05     ActorSelection
06     sel=system.actorSelection("akka://lifecycle/user/Supervisor/resta
07
08     for(int i=0;i<100;i++){
09         sel.tell(RestartActor.Msg.RESTART, ActorRef.noSender())

```

```

09     }
10 }
11 public static void main(String[] args) {
12     ActorSystem system = ActorSystem.create("lifecycle",
ConfigFactory.load("lifecycle.conf"));
13     customStrategy(system);
14 }

```

上述代码中，第12行代码创建了全局ActorSystem，接着在customStrategy()函数中创建了Supervisor Actor，并且对Supervisor发送一个RestartActor的Props（第3行，这个消息会使得Supervisor创建RestartActor）。

接着，选中RestartActor实例（第5行）。第7~9行，向这个RestartActor发送100条RESTART消息。这会使得RestartActor抛出NullPointerException。

执行上述代码，部分输出如下（由于输出太多，这里只截取重要的部分）：

```

01 preStart hashCode:7302437
02 meet NullPointerException,restart
03 preRestart hashCode:7302437
04 [ERROR] [lifecycle-akka.actor.default-dispatcher-3] [akka://li
restartActor] null
05 java.lang.NullPointerException
06     at geym.akka.demo.lifecycle.RestartActor.onReceive(Restart
07     at akka.actor.UntypedActor$$anonfun$receive$1.applyOrElse(

```

```
08     at akka.actor.Actor$class.aroundReceive(Actor.scala:465)
09     at akka.actor.UntypedActor.aroundReceive(UntypedActor.scala:
10     at akka.actor.ActorCell.receiveMessage(ActorCell.scala:516)
11     at akka.actor.ActorCell.invoke(ActorCell.scala:487)
12     at akka.dispatch.Mailbox.processMailbox(Mailbox.scala:254)
13     at akka.dispatch.Mailbox.run(Mailbox.scala:221)
14     at akka.dispatch.Mailbox.exec(Mailbox.scala:231)
15     at scala.concurrent.forkjoin.ForkJoinTask.doExec(ForkJoinTask
16     at scala.concurrent.forkjoin.ForkJoinPool$WorkQueue.runTask
17     at scala.concurrent.forkjoin.ForkJoinPool.runWorker(ForkJoinPool
18     at scala.concurrent.forkjoin.ForkJoinWorkerThread.run(ForkJoinWorkerThread
19
20 preStart hashCode:23269863
21 postRestart hashCode:23269863
22 meet NullPointerException, restart
23 preRestart hashCode:23269863
24 preStart hashCode:24918371
25 postRestart hashCode:24918371
26 meet NullPointerException, restart
27 preRestart hashCode:24918371
28 preStart hashCode:12844205
29 postRestart hashCode:12844205
30 [ERROR] [lifecycle-akka.actor.default-dispatcher-2]
[akka://lifecycle/user/Supervisor/restartActor] null
31 meet NullPointerException, restart
32 .....
33 postStop hashCode:12844205
```

第1行的preStart表示RestartActor正在初始化，注意它的HashCode为7302437。接着，这个Actor遇到了NullPointerException。根据自定义的策略，这将导致它重启，因此，这就有了第3行的preRestart，因为preRestart在正式重启之前调用，因此HashCode还是7302437，表示当前Actor和上一个Actor还是同一个实例。接着，第4~19行打印了异常信息。

第20行进入了preStart()方法，它的HashCode为23269863。这说明系统已经为这个RestartActor生成了一个新的实例，原有的实例因为重启而被回收。新的实例将代替原有实例继续工作。这说明同一个RestartActor在系统的工作始终，未必能保持同一个实例。重启完成后，调用postRestart()方法（第21行）。实际上，Actor重启后的preStart()方法，就是在postRestart()中调用的（Actor父类的postRestart()会调用preStart()方法）。

在经过3次重启后，超过了监督策略中的单位时间内的重试上限。因此，系统不会再进行尝试，而是直接关闭RestartActor。上述输出中第33行就显示了这个过程，在最后一个RestartActor实例上，执行了停止方法。

7.6 选择Actor

在一个ActorSystem中，可能存在大量的Actor。如何才能有效地对大量Actor进行批量的管理和通信呢？Akka为我们提供了一个ActorSelection类，用来批量进行消息发送。限于篇幅有限，这里不再给出完整的代码，示意代码如下：

```
1 for(int i=0;i<WORDER_COUNT;i++){
2     workers.add(system.actorOf(Props.create(MyWorker.class,i),
3 }
4
5 ActorSelection selection = getContext().actorSelection("/user/w
6 selection.tell(5, getSelf());
```

上述代码第1~3行，批量生成了大量Actor。接着，我们要给这些worker发送消息，通过actorSelection()方法提供的选择通配符（第5行），可以得到代表所有满足条件的ActorSelection。第6行，通过这个ActorSelection实例，便可以向所有woker Actor发送消息。

7.7 消息收件箱（Inbox）

我们已经知道，所有Actor之间的通信都是通过消息来进行的。这是否意味着我们必须构建一个Actor来控制整个系统呢？答案是否定的，我们并不一定要这么做，Akka框架已经为我们准备了一个叫做“收件箱”的组件，使用收件箱，可以很方便地对Actor进行消息发送和接收，大大方便了应用程序与Actor之间的交互。

下面定义了当前示例中唯一一个Actor：

```
01 public class MyWorker extends UntypedActor {
02     private final LoggingAdapter log = Logging.getLogger(getCo
03     public static enum Msg {
04         WORKING, DONE, CLOSE;
05     }
06
07     @Override
08     public void onReceive(Object msg) {
09         if (msg == Msg.WORKING) {
10             log.info("I am working");
11         }
12         if (msg == Msg.DONE) {
13             log.info("Stop working");
14         } if (msg == Msg.CLOSE) {
15             log.info("I will shutdown");
16             getSender().tell(Msg.CLOSE, getSelf());
17         }
18     }
19 }
```

```

17         getContext().stop(getSelf());
18     } else
19         unhandled(msg);
20 }
21 }

```

上述代码中，MyWorker会根据收到的消息打印自己的工作状态。当接收到CLOSE消息时（第14行），会关闭自己，结束运行。

而在本例中，与这个MyWorker Actor交互的，并不是一个Actor，而是一个邮箱，邮箱的使用很简单：

```

01 public static void main(String[] args) {
02     ActorSystem system = ActorSystem.create("inboxdemo", ConfigFac
03     ActorRef worker = system.actorOf(Props.create(MyWorker.class
04
05     final Inbox inbox = Inbox.create(system);
06     inbox.watch(worker);
07     inbox.send(worker, MyWorker.Msg.WORKING);
08     inbox.send(worker, MyWorker.Msg.DONE);
09     inbox.send(worker, MyWorker.Msg.CLOSE);
10
11     while(true){
12         Object msg = inbox.receive(Duration.create(1, TimeUnit.S
13         if(msg==MyWorker.Msg.CLOSE){
14             System.out.println("My worker is Closing");
15         }else if(msg instanceof Terminated){
16             System.out.println("My worker is dead");

```

```
17         system.shutdown();
18         break;
19     }else{
20         System.out.println(msg);
21     }
22 }
23 }
```

上述代码中，第5行，根据ActorSystem构造了一个与之绑定的邮箱Inbox。接着使用邮箱监视MyWorker（第6行），这样就能在MyWorker停止后得到一个消息通知。第7~9行，通过邮箱向MyWorker发送消息。

在第11~21行，进行消息接收，如果发现MyWorker已经停止工作，则关闭整个ActorSystem（第17行）。

执行上述代码，输出如下（为节省版面，我对输出进行了一些简单的删减）：

```
[INFO] [inboxdemo-akka.actor.default-dispatcher-3] [akka://inboxd
working
[INFO] [inboxdemo-akka.actor.default-dispatcher-3] [akka://inboxd
working
[INFO] [inboxdemo-akka.actor.default-dispatcher-3] [akka://inboxd
shutdown
My worker is Closing
My worker is dead
```

上述输出的前3行为MyWorker的输出日志，表示MyWorker Actor的工作状态。后两行为主函数main()中对MyWorker消息的处理。

7.8 消息路由

Akka提供了非常灵活的消息发送机制。有时候，我们也许会使用一组Actor而不是一个Actor来提供一项服务。这一组Actor中所有的Actor都是对等的，也就是说你可以找任何一个Actor来为你服务。这种情况下，如何才能快速有效地找到合适的Actor呢？或者说如何调度这些消息，才可以使负载更为均衡地分配在这一组Actor中。

为了解决这个问题，Akka使用一个路由器组件（Router）来封装消息的调度。系统提供了几种实用的消息路由策略，比如，轮询选择Actor进行消息发送，随机消息发送，将消息发送给最为空闲的Actor，甚至是在组内广播消息。

下面就来演示一下消息路由的使用方式：

```
01 public class WatchActor extends UntypedActor {
02     private final LoggingAdapter log = Logging.getLogger(getCo
03     public Router router;
04     {
05         List<Routee> routees=new ArrayList<Routee>();
06         for(int i=0;i<5;i++){
07             ActorRef worker = getContext().actorOf(Props.create
08             getContext().watch(worker);
09             routees.add(new ActorRefRoutee(worker));
10         }
11         router=new Router(new RoundRobinRoutingLogic(),routees
```

```

12     }
13
14     @Override
15     public void onReceive(Object msg) {
16         if(msg instanceof MyWorker.Msg){
17             router.route(msg, getSender());
18         }else if (msg instanceof Terminated) {
19             router=router.removeRoutee(((Terminated)msg).actor());
20             System.out.println(((Terminated)msg).actor().path()+"
routes().size());
21             if(router.routes().size()==0){
22                 System.out.println("Close system");
23                 RouteMain.flag.send(false);
24                 getContext().system().shutdown();
25             }
26         } else {
27             unhandled(msg);
28         }
29     }
30 }

```

上述代码中定义了WatchActor。第3行，就是路由器组件Router，在构造Router时，需要指定路由策略和一组被路由的Actor（Routee），如第11行所示。这里使用了RoundRobinRoutingLogic路由策略，也就是对所有的Routee进行轮询消息发送。在本例中，Routee由5个MyWorker Actor构成（第6～10行，MyWorker与上一节中的相同，故不再给出代码）。

当有消息需要传递给这5个MyWorker时，只需要将消息投递给这个Router即可（上述代码第17行）。Router就会根据给定的消息路由策略进行消息投递。当一个MyWorker停止工作时，还可以简单地将其从工作组中移除（第19行）。在这里，如果发现系统中没有可用的Actor，就会直接关闭系统。

主函数比较简单，如下：

```
01 public class RouteMain {
02     public static Agent<Boolean> flag=Agent.create(true, Execut
03     public static void main(String[] args) throws InterruptedExc
04         ActorSystem system = ActorSystem.create("route", ConfigFac
05         ActorRef w=system.actorOf(Props.create(WatchActor.class),
06         int i=1;
07         while(flag.get()){
08             w.tell(MyWorker.Msg.WORKING, ActorRef.noSender());
09             if(i%10==0)w.tell(MyWorker.Msg.CLOSE, ActorRef.noSende
10             i++;
11             Thread.sleep(100);
12         }
13     }
14 }
```

上述代码向WatchActor发送大量消息，其中夹杂着几条关闭Actor的消息。这会使得MyWorker Actor逐一被关闭，最终程序将退出。

这段程序的部分输出如下（做过适量裁剪）：

```
[INFO][route-akka.actor.default-dispatcher-3] [akka://route/user/working
[INFO][route-akka.actor.default-dispatcher-3] [akka://route/user/working
[INFO][route-akka.actor.default-dispatcher-3] [akka://route/user/working
[INFO][route-akka.actor.default-dispatcher-4] [akka://route/user/working
[INFO][route-akka.actor.default-dispatcher-3] [akka://route/user/working
[INFO][route-akka.actor.default-dispatcher-3] [akka://route/user/working
...
[INFO][route-akka.actor.default-dispatcher-2] [akka://route/user/shutdown
akka://route/user/watcher/worker_1 is closed, routees=0
Close system
```

可以看到，WORKING消息被轮流发送给这5个worker。大家可以修改路由策略，观察不同路由策略下的消息投递方式（除了RoundRobinRoutingLogic外，还可以尝试BroadcastRoutingLogic广播策略、RandomRoutingLogic随机投递策略、SmallestMailboxRoutingLogic空闲Actor优先投递策略）。

7.9 Actor的内置状态转换

在很多场景下，Actor的业务逻辑可能比较复杂，Actor可能需要根据不同的状态对同一条消息作出不同的处理。Akka已经为我们考虑到了这一点，一个Actor内部消息处理函数可以拥有多个不同的状态，在特定的状态下，可以对同一消息进行不同的处理，状态之间也可以任意切换。

现在让我们模拟一个婴儿Actor，假设婴儿会拥有两种不同的状态，开心或者生气。当你带他玩的时候，他总是会表现出开心状态，当你让他睡觉时，他就会非常生气，小孩子总是拥有用不完的精力，入睡困难可能是一种通病吧！

在我们这个简单的场景模拟中，我们会给这个婴儿Actor发送睡觉和玩两种指令。如果婴儿正在生气，你还让他睡觉，他就会说“我已经生气了”，如果你让他去玩，他就会变得开心起来。同样，如果他正玩得高兴，你让他继续玩，他就会说“我已经很愉快了”，如果让他睡觉，他就马上变得很生气。

下面的这个BabyActor就模拟了上述场景：

```
01 public class BabyActor extends UntypedActor {
02     private final LoggingAdapter log = Logging.getLogger(getCo
03     public static enum Msg {
04         SLEEP, PLAY, CLOSE;
05     }
```

```
06
07     Procedure<Object> angry = new Procedure<Object>() {
08         @Override
09         public void apply(Object message) {
10             System.out.println("angryApply:"+message);
11             if (message == Msg.SLEEP) {
12                 getSender().tell("I am already angry", getSelf
13                 System.out.println("I am already angry");
14             } else if (message == Msg.PLAY) {
15                 System.out.println("I like playing");
16                 getContext().become(happy);
17             }
18         }
19     };
20
21     Procedure<Object> happy = new Procedure<Object>() {
22         @Override
23         public void apply(Object message) {
24             System.out.println("happyApply:"+message);
25             if (message == Msg.PLAY) {
26                 getSender().tell("I am already happy :-)", get
27                 System.out.println("I am already happy :-)");
28             } else if (message == Msg.SLEEP) {
29                 System.out.println("I don't want to sleep");
30                 getContext().become(angry);
31             }
32         }
33     }
```

```

33     };
34
35     @Override
36     public void onReceive(Object msg) {
37         System.out.println("onReceive:"+msg);
38         if (msg == Msg.SLEEP) {
39             getContext().become(angry);
40         } else if (msg == Msg.PLAY) {
41             getContext().become(happy);
42         } else {
43             unhandled(msg);
44         }
45     }
46 }

```

上述代码中，使用了**become()**方法用于切换Actor的状态（第39、41行）。方法**become()**接收一个**Procedure**参数。**Procedure**在这里可以表示一种Actor的状态，同时，更重要的是它封装了在这种状态下的消息处理逻辑。

在这个**BabyActor**中，定义了两种**Procedure**，一是**angry**生气（第7行），另一个是**happy**开心（第21行）。

在初始状态下，**BabyActor**既没有生气也不开心。因此**angry**处理函数和**happy**处理函数都不会工作。当**BabyActor**接收到消息时，系统会调用**onReceive()**方法来处理这个消息。

令人吃惊的魔法就在这个**onReceive()**函数中。当**onReceive()**处理

SLEEP消息时，它会切换当前Actor的状态为angry（第39行）。如果是PLAY消息，则切换状态为happy。

一旦完成状态切换，当后续有新的消息送达时，就不会再由onReceive()函数处理了。由于angry和happy本身就是消息处理函数。因此，后续的消息就直接交由当前状态处理（angry或者happy），从而很好地封装了Actor的多个不同处理逻辑。

下面的代码向我们的婴儿Actor发送了几条PLAY和SLEEP的消息：

```
1 ActorSystem system = ActorSystem.create("become", ConfigFactory
2 ActorRef child = system.actorOf(Props.create(BabyActor.class),
3 system.actorOf(Props.create(WatchActor.class, child), "watcher"
4 child.tell(BabyActor.Msg.PLAY, ActorRef.noSender());
5 child.tell(BabyActor.Msg.SLEEP, ActorRef.noSender());
6 child.tell(BabyActor.Msg.PLAY, ActorRef.noSender());
7 child.tell(BabyActor.Msg.PLAY, ActorRef.noSender());
8
9 child.tell(PoisonPill.getInstance(), ActorRef.noSender());
```

其输出如下（进行过适量裁剪）：

```
onReceive:PLAY
happyApply:SLEEP
I don't want to sleep
angryApply:PLAY
I like playing
happyApply:PLAY
```

```
I am already happy :-)  
[INFO][akka://become/user/watcher] akka://become/user/baby has te  
system
```

可以看到，当第一个PLAY消息到来时，是由onReceive()函数进行处理的，在onReceive()中，将Actor切换为happy状态。因此，当SLEEP消息达到时，由happy.apply()函数处理，接着Actor切换为angry状态。当PLAY消息再次到达时，由angry.apply()函数处理。由此可见，Akka为Actor提供了灵活的状态切换机制，处于不同状态的Actor可以绑定不同的消息处理函数进行消息处理，这对构造结构化应用有着重要的帮助。

7.10 询问模式：Actor中的Future

由于Actor之间都是通过异步消息通信的。当你发送一条消息给一个Actor后，你通常只能等待Actor的返回。与同步方法不同，在你发送异步消息后，接受消息的Actor可能还根本来不及处理你的消息，而调用方就已经返回了。

这种模式与我们之前提到的Future模式非常相像。不同之处只是在传统的异步调用中，我们进行的是函数调用，但在这里，我们发送了一条消息。

因为两者的行为方式是如此相像，因此我们就会很自然地想到，当我们需要一个有返回值的调用时，Actor是不是也应该给我们一个契约（Future）呢？这样，就算我们当下没有办法立即获得Actor的处理结果，在将来，通过这个契约还是可以追踪到我们的请求的。

```
01 import static akka.pattern.Patterns.ask;
02 import static akka.pattern.Patterns.pipe;
03
04 public class AskMain {
05
06     public static void main(String[] args) throws Exception {
07         ActorSystem system = ActorSystem.create("askdemo", Con
08         ActorRef worker = system.actorOf(Props.create(MyWorker
09         ActorRef printer = system.actorOf(Props.create(Printer
10         system.actorOf(Props.create(WatchActor.class, worker),
```

```
11
12     //等待future返回
13     Future<Object> f = ask(worker, 5, 1500);
14     int re = (int) Await.result(f, Duration.create(6, Time
15     System.out.println("return:" + re);
16
17     //直接导向其他Actor, pipe不会等待
18     f = ask(worker, 6, 1500);
19     pipe(f, system.dispatcher()).to(printer);
20
21     worker.tell(PoisonPill.getInstance(), ActorRef.noSende
22 }
23 }
```

上述代码给出了两处Actor交互中使用Future的例子。

在第13行，使用ask()方法给worker发送消息，消息内容是5，也就是说worker会接收到一个Integer消息，值为5。当worker接收到消息后，就可以进行计算处理，并且将结果返回给发送者。当然，这个处理过程可能需要花费一点时间。

方法ask()不会等待worker处理，会立即返回一个Future对象（第13行）。在第14行，我们使用Await方法等待worker的返回，接着在第15行打印返回结果。

在这种方法中，我们间接地将一个异步调用转为同步阻塞调用。虽然比较容易理解，但是在有些场合可能会出现性能问题。另外一种更为有效的方法是使用pipe()函数。

代码第18行使用ask()再次询问worker，并传递数值6给worker。接着并不进行等待，而是使用pipe()将这个Future重定向到另外一个称为printer的Actor。pipe()函数不会阻塞程序，会立即返回。

这个printer的实现很简单的，只是简单地输出得到的数据：

```
01 @Override
02 public void onReceive(Object msg) {
03     if (msg instanceof Integer) {
04         System.out.println("Printer:"+msg);
05     }
06     if (msg == Msg.DONE) {
07         log.info("Stop working");
08     } if (msg == Msg.CLOSE) {
09         log.info("I will shutdown");
10         getSender().tell(Msg.CLOSE, getSelf());
11         getContext().stop(getSelf());
12     } else
13         unhandled(msg);
14 }
```

上述代码就是Printer Actor的实现，它会通过pipe()方法得到worker的输出结果，并打印在控制台上（第4行）。

在本例中，worker Actor接受一个整数，并计算它的平方，并给予返回。如下：

```
01 @Override
```



```
02 public void onReceive(Object msg) {
03     if (msg instanceof Integer) {
04         int i=(Integer)msg;
05         try {
06             Thread.sleep(1000);
07         } catch (InterruptedException e) {}
08         getSender().tell(i*i, getSelf());
09     }
10     if (msg == Msg.DONE) {
11         log.info("Stop working");
12     }if (msg == Msg.CLOSE) {
13         log.info("I will shutdown");
14         getSender().tell(Msg.CLOSE, getSelf());
15         getContext().stop(getSelf());
16     } else
17         unhandled(msg);
18 }
```

上述代码第5~7行，模拟了一个耗时的调用，为了更明显地说明ask()和pipe()方法的用途。第8行，worker计算了给定数值的平方，并把它“告诉”请求者。

7.11 多个Actor同时修改数据： Agent

在Actor的编程模型中，Actor之间主要通过消息进行信息传递。因此，很少发生多个Actor需要访问同一个共享变量的情况。但在实际开发中，这种情况很难完全避免。那如果多个Agent需要对同一个共享变量进行读写时，如何保证线程安全呢？

在Akka中，使用一种叫做Agent的组件来实现这个功能。一个Agent提供了对一个变量的异步更新。当一个Actor希望改变Agent的值时，它会向这个Agent下发一个动作（action）。当多个Actor同时改变Agent时，这些action将会在执行上下文（ExecutionContext）中被并发调度执行。在任意时刻，一个Agent最多只能执行一个action，对于某一个线程来说，它执行action的顺序与它的发生顺序一致，但对于不同线程来说，这些action可能会交织在一起。

Agent的修改可以使用两个方法send()或者alter()。它们都可以向Agent发送一个修改动作。但是send()方法没有返回值，而alter()方法会返回一个Future对象便于跟踪Agent的执行。

下面让我们模拟这么一个场景：有10个Actor，它们一起对一个Agent执行累加操作，每个agent累加10000次，如果没有意外，那么agent最终的值将是100000，如果Actor间的调度出现问题，那么这个值可能小于100000。

```
01 public class CounterActor extends UntypedActor {
```

```

02     Mapper addMapper = new Mapper<Integer, Integer>() {
03         @Override
04         public Integer apply(Integer i) {
05             return i+1;
06         }
07     };
08
09     @Override
10     public void onReceive(Object msg) {
11         if (msg instanceof Integer) {
12             for (int i = 0; i < 10000; i++) {
13                 //我希望能够知道future何时结束
14                 Future<Integer> f = AgentDemo.counterAgent.alter(1, addMapper);
15                 AgentDemo.futures.add(f);
16             }
17             getContext().stop(getSelf());
18         } else
19             unhandled(msg);
20     }
21 }

```

上述代码定义了一个累加的Actor: CounterActor。第2~7行, 定义了累计动作action addMapper。它的作用是对Agent的值进行修改, 这里简单地加1。

CounterActor的消息处理函数onReceive()中, 对全局的counterAgent进行累加操作, alter()指定了累加动作addMapper (第14行)。由于我们

希望在将来知道累加行为是否完成，因此在这里将返回的Future对象进行收集（第15行）。完成任务后，Actor自行退出（第17行）。

程序的主函数如下：

```
01 public class AgentDemo {
02     public static Agent<Integer> counterAgent = Agent.create(
03         static ConcurrentLinkedQueue<Future<Integer>> futures =
<Integer>>());
04
05     public static void main(String[] args) throws InterruptedException
06         final ActorSystem system = ActorSystem.create("agentde
07             ConfigFactory.load("samplehello.conf"));
08         ActorRef[] counter = new ActorRef[10];
09         for (int i = 0; i < counter.length; i++) {
10             counter[i] = system.actorOf(Props.create(CounterAc
11         }
12         final Inbox inbox = Inbox.create(system);
13         for (int i = 0; i < counter.length; i++) {
14             inbox.send(counter[i], 1);
15             inbox.watch(counter[i]);
16         }
17
18         int closeCount = 0;
19         //等待所有Actor全部结束
20         while (true) {
21             Object msg = inbox.receive(Duration.create(1, Time
```

```

22         if (msg instanceof Terminated) {
23             closeCount++;
24             if (closeCount == counter.length) {
25                 break;
26             }
27         } else {
28             System.out.println(msg);
29         }
30     }
31     // 等待所有的累加线程完成,因为他们都是异步的
32     Futures.sequence(futures, system.dispatcher()).onComple
33         new OnComplete<Iterable<Integer>>() {
34             @Override
35             public void onComplete(Throwable arg0, It
36                 System.out.println("counterAgent=" + c
37                 system.shutdown();
38             }
39         }, system.dispatcher());
40 }
41 }

```

上述代码中，第8~11行，创建了10个CounterActor对象。第12~16行，使用Inbox与CounterActor进行通信。第14行的消息将触发CounterActor进行累加操作。第20~30行系统将等待所有10个CounterActor运行结束。执行完成后，我们便已经收集了所有的Future。在第32行，将所有的Future进行串行组合（使用sequence()方法），构造了一个整体的Future，并为它创建onComplete()回调函数。在所有的

Agent操作执行完成后，onComplete()方法就会被调用（第35行）。在这个例子中，我们简单地输出最终的counterAgent值（第36行），并关闭系统（第37行）。

执行上述程序，我们将看到：

```
counterAgent=100000
```

7.12 像数据库一样操作内存数据：软件事务内存

在一些函数式编程语言中，支持一种叫做软件事务内存（STM）的技术。什么是软件事务内存呢？这里的事务和数据库中所说的事务非常类似，具有隔离性、原子性和一致性。与数据库事务不同的是，内存事务不具备持久性（很显然内存数据不会保存下来）。

在很多场合，某一项工作可能要由多个Actor协作完成。在这种协作事务中，如果一个Actor处理失败，那么根据事务的原子性，其他Actor所进行的操作必须要回滚。下面，就让我们来看一个简单的案例。

假设有一个公司要给他的员工发放福利，公司账户里有100元。每次，公司账户会给员工账户转一笔钱，假设转账10元，那么公司账户中应该减去10元，同时，员工账户中应该增加10元。这两个操作必须同时完成，或者同时不完成。

首先，让我们看一下主函数中是如何启动一个内存事务的：

```
01 public class STMDemo {
02     public static ActorRef company=null;
03     public static ActorRef employee=null;
04
05     public static void main(String[] args) throws Exception {
06 final ActorSystem system = ActorSystem.create("transactionDemo
```

```

("samplehello.conf"));
07         company=system.actorOf(Props.create(CompanyActor.class
08         employee=system.actorOf(Props.create(EmployeeActor.cla
09
10         Timeout timeout = new Timeout(1, TimeUnit.SECONDS);
11
12         for(int i=1;i<20;i++){
13             company.tell(new Coordinated(i, timeout), ActorRef
14             Thread.sleep(200);
15             Integer companyCount = (Integer) Await.result(
16                 ask(company, "GetCount", timeout), timeout
17             Integer employeeCount = (Integer) Await.result(
18                 ask(employee, "GetCount", timeout), timeou
19
20             System.out.println("company count="+companyCount);
21             System.out.println("employee count="+employeeCount
22             System.out.println("=====");
23         }
24     }
25 }

```

上述代码中CompanyActor和EmployeeActor分别用于管理公司账户和雇员账户。在第12~23行中，我们尝试进行19次汇款，第一次汇款额度为1元，第二次为2元，依此类推，最后一笔汇款为19元。

在第13行，新建一个Coordinated协调者，并且将这个协调者当做消息发送给company。当company收到这个协调者消息后，自动成为这个

事务的第一个成员。

第15~18行询问公司账户和雇员账户的当前余额，并在第20~21行进行输出。

下面是代表公司账户的Actor:

```
01 public class CompanyActor extends UntypedActor {
02     private Ref.View<Integer> count = STM.newRef(100);
03
04     @Override
05     public void onReceive(Object msg) {
06         if (msg instanceof Coordinated) {
07             final Coordinated c=(Coordinated)msg;
08             final int downCount=(Integer)c.getMessage();
09             STMDemo.employee.tell(c.coordinate(downCount), get
10             try{
11                 c.atomic(new Runnable() {
12                     @Override
13                     public void run() {
14                         if(count.get()<downCount){
15                             throw new RuntimeException("less t
16                         }
17                         STM.increment(count, -downCount);
18                     }
19                 });
20             }catch(Exception e){
21                 e.printStackTrace();
```

```

22         }
23
24         }else if ("GetCount".equals(msg)) {
25             getSender().tell(count.get(), getSelf());
26         }else{
27             unhandled(msg);
28         }
29     }
30 }

```

在CompanyActor中，首先判断接收的msg是否是Coordinated。如果是Coordinated，则表示这是一个新事务的开始。在第8行，获得事务的参数也就是需要转账的金额。接着在第9行，将调用Coordinated.coordinate()方法，将employee也加入到当前事务中，这样这个事务中就有两个参与者了。

第11行，调用了Coordinated.atomic()定义了原子执行块作为这个事务的一部分。在这个执行块中，对公司账户进行余额调整（第17行）。但是当汇款额度大于可用余额时，就会抛出异常，宣告失败。

第25行用于处理GetCount消息，返回当前账户余额。

作为转账接收方的雇员账户如下：

```

01 public class EmployeeActor extends UntypedActor {
02     private Ref.View<Integer> count = STM.newRef(50);
03
04     @Override

```

```

05     public void onReceive(Object msg) {
06         if (msg instanceof Coordinated) {
07             final Coordinated c = (Coordinated) msg;
08             final int downCount = (Integer) c.getMessage();
09             try {
10                 c.atomic(new Runnable() {
11                     @Override
12                     public void run() {
13                         STM.increment(count, downCount);
14                     }
15                 });
16             } catch (Exception e) {
17             }
18         } else if ("GetCount".equals(msg)) {
19             getSender().tell(count.get(), getSelf());
20         } else {
21             unhandled(msg);
22         }
23     }
24 }

```

上述代码第2行，设置雇员账户初始金额是50元。第6行，判断消息是否为Coordinated，如果是Coordinated，则当前Actor会自动加入Coordinated指定的事务。第10行，定义原子操作，在这个操作中将修改雇员账户余额。在这里，我们并没有给出异常情况的判断，只要接收到转入金额，一律将其增加到雇员账户中。

大家可能就会产生疑问，如果在公司账户中由于余额不足而导致转账失败了，那在这个雇员账户中不还是正常增加了金额吗？那岂不是钱多出来了？

不过这个担心是完全多余的。因为在这里，两个Actor都已经加入到同一个协调事务Coordinated中了，因此当公司账户出现异常后，雇员账户的余额就会回滚。

执行上述程序，部分输出如下：

```
.....
company count=85
employee count=65
=====
java.lang.RuntimeException: less than 14
company count=9
employee count=141
....
=====
java.lang.RuntimeException: less than 19
    省略堆栈信息 实在太多了
    at scala.concurrent.forkjoin.ForkJoinWorkerThread.run(ForkJoi
company count=9
employee count=141
=====
```

可以看到，无论转账操作是否成功，公司账户和雇员账户的金额总是一致的。当转账失败时，雇员账户的余额并不会增加。这就是软件事

务内存的作用。

7.13 一个有趣的例子：并发粒子群的实现

粒子群算法（PSO）是一种进化算法。它与大名鼎鼎的遗传算法非常类似，可以用来解决一些优化问题。大家知道，一些优化问题（比如旅行商问题TSP）都属于NP问题。它们的时间复杂度可能会达到 $O(n!)$ 或者 $O(2^n)$ ，这种在多项式时间内不可解的问题总是会让人望而生畏。而以PSO算法为代表的进化计算，往往可以将这些NP问题，转变为一个多项式问题。但这种转变是有代价的，进化算法往往都不保证你可以从结果中得到最优解。我这么说，也许就有人会问了，这个算法都不能保证得到最优解，那有什么用呢？其实，在生活中的很多场景下，并不是特别需要最优解，我们更加希望得到的是一个满意解。比如说，去水果店买西瓜，店里可能放着一大堆西瓜，每个人都想挑一个最好的。但你想拿到最好的那个西瓜必须得挨个检查过去，并且还得认真做好记录才行。我相信，没有一个人会这么买西瓜，因为成本太高了。对于大部分人来说，更倾向于在表面上挑几个顺眼的看看，如果还过得去，也就下手了。这也就是说只要这个结果不要差得太离谱就行了。

既然最优的方案很难得到，那么我们就想办法以很低的成本获得一个还算过得去的方案，也不失为一计良策。在后面给出的小案例中大家也可以看到，在很多情况下，虽然进化算法无法让你获得最优解，也无法证明它得到的解与最优解到底有多少差距，但实际中，通过进化算法搜索到的满意解很可能与最优解已经非常接近了。

7.13.1 什么是粒子群算法

粒子群优化算法（PSO）是一种进化计算技术，最早由Kenny与Eberhart于1995年提出。它源于对鸟群捕食行为的研究，与遗传算法相似，是一种基于迭代的优化算法，广泛应用于函数优化和神经网络训练等方面。与遗传算法相比，PSO算法的实现简单得多，参数配置也相对较少，对使用人员的经验要求不高，因此更加易于实际工程应用。

从日常生活的观察中可以知道，鸟类的觅食往往会表现成群体特性。如果在地上有一小撮食物，那么鸟群很可能就会聚集在这一堆食物旁边。如果其中一只小鸟发现了另外一堆更丰盛的食物，那它可能会离群飞向更丰盛的食物，而这有可能带动整个鸟群一起飞向新的地点。当然了，在整个种群中，难免会出现几只特别有“个性”的小鸟，它们不喜欢太热闹的地方，当整个种群迁移时，它们不会跟着种群走，或者自己散步，或者自行游荡。

粒子群算法正是对上述过程的模拟。在程序中，我们可以模拟大量的小鸟，小鸟的觅食点正是要求解的问题的解。解越是优秀，意味着食物越是丰盛，因此，模拟的小鸟会从自己的位置出发以一定的速度向最优点的方向移动。在移动过程中，任何一只小鸟都有可能发现更好的解，这又会进一步影响群体的行为。就这样如此反复迭代，最终，将得到一个不错的答案。

7.13.2 粒子群算法的计算过程

粒子群算法的大体步骤如下：

1. 初始化所有粒子，粒子的位置随机生成。计算每个粒子当前的适应度，并将此设为当前粒子的个体最优值（记为pBest）。
2. 所有粒子将自己的个体最优值发送给管理者Master。Master获得所有粒子的信息后，筛选出全局最优的解（记为gBest）。
3. Master将gBest通知所有粒子，所有粒子便知道全局最优点的位置。
4. 接着，所有粒子根据自己的pBest和全局gBest，更新自己的速度，在有了速度后，再更新自己的位置。

$$v_{k+1}=c_0 \times \text{rand}() \times v_k + c_1 \times \text{rand}() \times (\text{pbest}_k - x_k) + c_2 \times \text{rand}() \times (\text{gbest}_k - x_k)$$

$$x_{k+1}=x_k + v_{k+1}$$

其中，rand()函数产生一个0，1之间的随即数。 $c_0=1$ ， $c_1=2$ ，

$c_2=2$ ，k表示进化的代数。 v_k 表示当前速度， pbest_k 和 gbest_k 表示个体最优解和全局最优解。当然，对于每一个维度上的速度分量，我们可以为它限定一个最大值。确保“小鸟”不会飞得太快，错过了重要的信息。

5. 如果粒子产生了新的个体最优点，则发送给Master，在此，转到步骤2。

整体过程的示意图如图7.2所示。

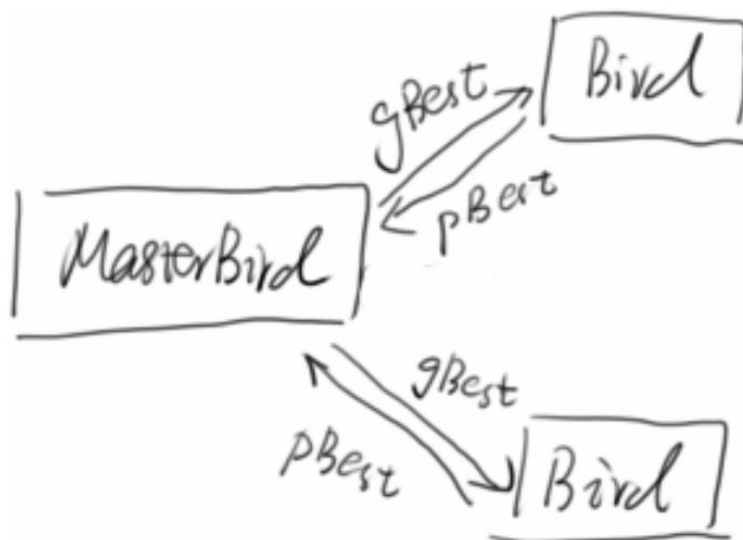


图7-2 PSO算法示意图

从这个计算步骤中可以看到，计算过程拥有一定的随机性。但由于我们可以启用大量的例子，因此其计算效果在统计学意义上是稳定的。在这个标准的粒子群算法中，由于所有粒子都会向全局最优靠拢，因此，其跳出局部最优的能力并不算太强。因此，我们也可以想办法对标准的粒子群算法进行一些合理的改进。比如，允许各个粒子随机移动，甚至逆向移动来试图突破局部最优。在这里为简单起见，我不打算做这些复杂的实现。

7.13.3 粒子群算法能做什么

粒子群算法能为我们做些什么呢？它应用最多的场景是进行最优化计算。实际上，以粒子群算法为代表的进化计算，可以说是最优化方法中的通用方法。几乎一切最优化问题都可以通过这种随机搜索的模式解决，其成本低、难度小、效果好，因此颇受欢迎。

下面，就让我们来探讨一个典型的优化问题：

假设现在有400万资金，要求4年内使用完。若在第1年使用 x 万元，则可以得到效益 \sqrt{x} 万元（效益不能再使用），当年不用的资金可存入银行，年利率为10%。尝试制订出资金的使用规划，使4年效益之和最大。

很明显，对于这类问题，不同的方案得到的结果可能会有很大的差异。比如，若第一年把400万元全部用完，则总效益为 $\sqrt{400} = 20$ 万元；若前3年均不用而存入银行，第4年把本金和利息全部用完，则总效益为万元，显然优于第一种方案。

如果我们将此问题转为一般化的优化问题，则可以得到以下方程组，如图7.3所示。

$$\begin{aligned} \max \quad & Z = \sqrt{x_1} + \sqrt{x_2} + \sqrt{x_3} + \sqrt{x_4} \\ \text{s.t.} \quad & x_1 \leq 400 \\ & 1.1x_1 + x_2 \leq 440 \\ & 1.21x_1 + 1.1x_2 + x_3 \leq 484 \\ & 1.331x_1 + 1.21x_2 + 1.1x_3 + x_4 \leq 532.4 \\ & x_1, x_2, x_3, x_4 \geq 0 \end{aligned}$$

图7.3 一般化的约束问题

其中， x_1 、 x_2 、 x_3 、 x_4 分别表示第1、2、3、4年使用的资金。使用拉格朗日乘子法对此方程组进行求解，可以得到第一年使用86.19万元、第2年使用104.29万元，第3年使用126.19万元，第4年使用152.69万元为这个问题的最优解，此时总效益达43.09万元。

由于求解过程过于复杂，使用拉格朗日乘子法时，需要对先后12个

未知数和方程进行联立求解，比较难以实现。由于求解过程与我们讨论的主题无关，所以在这里不再给出。

对于类似的优化问题，正是粒子群算法的涉猎范围。当使用粒子群算法时，我们可以先随机给出若干个满足提交的资金规划方案。接着，根据粒子群的演化公式，不断调整各个粒子的位置（粒子的每一个位置代表一套方案），逐步探索更优的方案。

7.13.4 使用Akka实现粒子群

现在，我们已经知道粒子群的原理，并且有了一个较为复杂的优化问题等待我们求解。接下来，就需要开动脑筋，使用Akka来实现一个简单的粒子群，来解决这个优化问题了。

使用Actor的模式与粒子群算法之间有着天生切合度。粒子群算法由于涉及到多个甚至是极其大量的粒子参与运算，因此它隐含着并行计算的模式。其次，从直观上我们也可以知道，粒子群算法的求解精度或者说求解的质量，与参与运算的粒子有着直接的关系。很显然，参与运算的粒子数量越多，得到的解自然也就越精确。

如果我们使用传统的多线程方式实现粒子群，一个最大的问题就是线程的数量可能是非常有限的。在当前这种应用场景中，我们希望可以拥有数万，甚至数十万的粒子，以提高计算精度，但众所周知，在一台计算机上运行数万个线程基本是不可能的，就算可以，系统的性能也会大打折扣。因此，使用多线程的模型无法很好地和粒子群的实现相融合。

但Akka的Actor的模型则不同。由于多个Actor可以复用一条线程，而Actor本身作为轻量级的并发执行单元可以有极其大量的存在。因此，我们就可以使用Actor来模拟整个粒子群计算的场景。下面就让我们仔细看一下系统的实现。

首先，我们需要两个表示pBest和gBest的消息类型，用于在多个Actor之间传递个体最优和全局最优。

```
01 public final class GBestMsg {
02     final PsoValue value;
03     public GBestMsg(PsoValue v){
04         value=v;
05     }
06     public PsoValue getValue() {
07         return value;
08     }
09 }
10
11 public final class PBestMsg {
12     final PsoValue value;
13     public PBestMsg(PsoValue v){
14         value=v;
15     }
16
17     public PsoValue getValue() {
18         return value;
19     }
}
```

```

20
21     public String toString(){
22         return value.toString();
23     }
24 }

```

上述代码中，GBestMsg（代码第1行）表示携带全局最优解的消息。而PBestMsg（代码第11行）表示携带个体最优的消息。它们都使用PsoValue来表示一个可行的解。

在PsoValue中，主要包括两个信息，第一是表示投资规划的方案，即每一年分别需要投资多少钱；第二是这个投资方案的总收益：

```

01 public final class PsoValue {
02     final double value;
03     final List<Double> x;
04     public PsoValue(double v,List<Double> x){
05         value=v;
06         List<Double> b=new ArrayList<Double>(5);
07         b.addAll(x);
08         this.x=Collections.unmodifiableList(b);
09     }
10     public double getValue(){
11         return value;
12     }
13     public List<Double> getX(){
14         return x;
15     }

```

```

16
17     public String toString(){
18         StringBuffer sb=new StringBuffer();
19         sb.append("value:").append(value).append("\n")
20         .append(x.toString());
21         return sb.toString();
22     }
23 }

```

上述代码中，数组x中，x[1]、x[2]、x[3]、x[4]分别表示第1年、第2年、第3年和第4年的投资额。这里为了方便起见，我忽略了x[0]（它在我们的程序中是没有作用的）。成员变量value表示这组投资方案的收益值。

因此，根据需求x与value之间的关系如下代码所示：

```

1 public class Fitness {
2     public static double fitness(List<Double> x){
3         double sum=0;
4         for(int i=1;i<x.size();i++){
5             sum+=Math.sqrt(x.get(i));
6         }
7         return sum;
8     }
9 }

```

上述代码定义的fitness()函数返回了给定投资方案的适应度。适应度也就是投资的收益，我们自然应该更倾向于选择适应度更高的投资方

案。在这里适应度= $\sqrt{x1} + \sqrt{x2} + \sqrt{x3} + \sqrt{x4}$ 。

有了这些基础工具，我们就可以来实现简单的粒子（这里我把它叫作Bird）了。

对于基本粒子，我们需要定义以下成员变量：

```
1 public class Bird extends UntypedActor {
2     private final LoggingAdapter log = Logging.getLogger(getCon
3     private PsoValue pBest=null;
4     private PsoValue gBest=null;
5     private List<Double> velocity =new ArrayList<Double>(5);
6     private List<Double> x =new ArrayList<Double>(5);
7     private Random r = new Random();
```

上述代码中，pBest和gBest分别表示个体最优和全局最优，velocity表示粒子在各个维度上的速度（在当前案例中，每一年的投资额就可以认为是一个维度，因此系统有4个维度）。x表示投资方案，即每一年的投资额。由于在粒子群算法中，需要使用随机数，因此，这里定义了r。

当一个粒子被创建时，我们需要初始化粒子的当前位置。粒子的每一个位置都代表一个投资方案，下面的代码展示了粒子的初始化逻辑：

```
01 @Override
02 public void preStart(){
03     for(int i=0;i<5;i++){
04         velocity.add(Double.NEGATIVE_INFINITY);
05         x.add(Double.NEGATIVE_INFINITY);
```

```

06     }
07     //x1<=400
08     x.set(1, (double)r.nextInt(401));
09
10     //x2<=440-1.1*x1
11     double max=400-1.1*x.get(1);
12     if(max<0)max=0;
13     x.set(2, r.nextDouble()*max);
14
15     //x3<=484-1.21*x1-1.1*x2
16     max=484-1.21*x.get(1)-1.1*x.get(2);
17     if(max<=0)max=0;
18     x.set(3, r.nextDouble()*max);
19
20     //x4<= 532.4-1.331*x1-1.21*x2-1.1*x3
21     max=532.4-1.331*x.get(1)-1.21*x.get(2)-1.1*x.get(3);
22     if(max<=0)max=0;
23     x.set(4, r.nextDouble()*max);
24
25     double newFit=Fitness.fitness(x);
26     pBest=new PsoValue(newFit,x);
27     PBestMsg pBestMsg=new PBestMsg(pBest);
28     ActorSelection selection = getContext().actorSelection("/u
29     selection.tell(pBestMsg, getSelf());
30 }

```

由于在当前案例中，每一年的投资额度是有条件约束的，比如第一

年的投资额不能超过400万（第7~8行），而第2年的投资上限是440万（假设第一年全部存银行，代码第10~13行），依此类推。粒子初始化时，随机生成一组满足基本约束条件的投资组合，并计算它的适应度（第25行）。初始的投资方案自然也就作为当前的个体最优，并发送给Master（第29行）。

当Master计算出当前全局最优后，会将全局最优发送给每一个粒子，粒子根据全局最优更新自己的运行速度，并更新自己的速度以及当前位置。

```
01 @Override
02 public void onReceive(Object msg) {
03     if (msg instanceof GBestMsg) {
04         gBest=((GBestMsg) msg).getValue();
05         //更新速度
06         for(int i=1;i<velocity.size();i++){
07             updateVelocity(i);
08         }
09         //更新位置
10         for(int i=1;i<x.size();i++){
11             updateX(i);
12         }
13         validateX();
14         double newFit=Fitness.fitness(x);
15         if(newFit>pBest.value){
16             pBest=new PsoValue(newFit,x);
17             PBestMsg pBestMsg=new PBestMsg(pBest);
```

```

18         getSender().tell(pBestMsg, getSelf());
19     }
20 }
21 else{
22     unhandled(msg);
23 }
24 }

```

上述代码中，粒子接收到了全局最优（代码第4行），接着根据粒子群的标准公式更新自己的速度（第6~8行）。接着，根据速度，更新自己的位置（第10~12行）。由于当前问题是有约束的，也就是说解空间并不是随意的。粒子很可能在更新位置后，跑出了合理的范围之外，因此，还有必要进行有效性检查（第13行）。

在更新完成后，就可以计算新位置的适应度，如果产生了新的个体最优，就将其发送给Master（第15~19行）。

在当前案例中，速度和位置的更新是依据标准的粒子群实现，如下：

```

01 public double updateVelocity(int i){
02     double v= Math.random()*velocity.get(i)
03         +2*Math.random()*(pBest.getX().get(i)-x.get(i))
04         +2*Math.random()*(gBest.getX().get(i)-x.get(i));
05     v=v>0? Math.min(v, 5): Math.max(v, -5);
06     velocity.set(i, v);
07     return v;
08 }

```

```
09
10 public double updateX(int i){
11     double newX=x.get(i)+velocity.get(i);
12     x.set(i, newX);
13     return newX;
14 }
```

上述代码中updateVelocity()和updateX()分别更新了粒子的速度和位置。位置的更新依赖于当前的速度（第11行）。

由于每一年的投资都是有限额的，因此，要避免粒子跑到合理空间之外，下面的代码强制将粒子约束中合理的区间中。

```
01 public void validateX(){
02     if(x.get(1)>400){
03         x.set(1, (double)r.nextInt(401));
04     }
05
06     //x2
07     double max=400-1.1*x.get(1);
08     if(x.get(2)>max || x.get(2)<0){
09         x.set(2, r.nextDouble()*max);
10     }
11     //x3
12     max=484-1.21*x.get(1)-1.1*x.get(2);
13     if(x.get(3)>max || x.get(3)<0){
14         x.set(3, r.nextDouble()*max);
15     }
```

```

16      //x4
17      max=532.4-1.331*x.get(1)-1.21*x.get(2)-1.1*x.get(3);
18      if(x.get(4)>max || x.get(4)<0){
19          x.set(4, r.nextDouble()*max);
20      }
21 }

```

上述代码分别对x1、x2、x3、x4进行约束，一旦发现粒子跑出了定义范围就将它进行随机化。

此外，我们还需要一只MasterBird，用于管理和通知全局最优。

```

01 public class MasterBird extends UntypedActor {
02     private final LoggingAdapter log = Logging.getLogger(getCo
03     private PsoValue gBest=null;
04
05     @Override
06     public void onReceive(Object msg) {
07         if (msg instanceof PBestMsg) {
08             PsoValue pBest = ((PBestMsg) msg).getValue();
09             if(gBest==null || gBest.value < pBest.value){
10                 //更新全局最优，通知所有粒子
11                 System.out.println(msg+"\n");
12                 gBest=pBest;
13                 ActorSelection selection = getContext().actors
14                 selection.tell(new GBestMsg(gBest), getSelf())
15             }
16         }

```

```
17         else{
18             unhandled(msg);
19         }
20     }
21 }
```

上述代码定义了MasterBird。当它收到一个个体最优的解时，会将其与全局最优进行比较，如果产生了新的全局最优，就更新这个全局最优并通知所有的粒子（第12~14行）。

好了，现在万事俱备只欠东风。下面就是主函数：

```
01 public class PSOMain {
02     public static final int BIRD_COUNT = 100000;
03     public static void main(String[] args) {
04         ActorSystem system = ActorSystem
05             .create("psoSystem", ConfigFactory.load("sampl
06         system.actorOf(Props.create(MasterBird.class), "master
07         for (int i = 0; i < BIRD_COUNT; i++) {
08             system.actorOf(Props.create(Bird.class), "bird_" +
09         }
10     }
11 }
```

上述代码定义了粒子总数，这里是10万个粒子。接着创建一个MasterBird Actor（第6行），和10万个bird（第7~9行）。

执行上述代码，运行一小段时间，你就可以得到如下输出（截取部

分)：

```
value:36.15412875487459
[-Infinity, 168.0, 18.786423873345715, 102.1742923174793, 76.5657
value:37.88452477135976
[-Infinity, 64.0, 87.66774733441137, 37.976681047619195, 206.1779
....
value:42.240797528048176
[-Infinity, 113.0, 42.37168995110633, 141.70570102409184, 174.168
....
value:43.01934824083668
[-Infinity, 76.0, 112.89557345993592, 133.29270155682005, 147.162
```

上述输出表示，当粒子群随机初始化时，最优解为36.15万元，但随着粒子的搜索，这个投资方案被逐步优化，由37.88万一直上升到43.02万元。根据我们前面的求解，我们知道这个投资方案的最优结果是43.09万元，可以看到，粒子群的搜索结果和全局最优已经非常接近了。

当然了，由于粒子群算法的随机性，每次执行结果可能并不一样，这意味着有时候，你可能会求得更好的解，或者得到一个稍差一些的解，但其偏差不会相差太远。

7.14 参考文献

- Akka官方文档
 - <http://doc.akka.io/docs/akka/2.3.7/java.html>
- 有关最优化方法的介绍
 - 《最优化方法》高等教育出版社施光燕著
- Nobody Needs Reliable Messaging
 - <http://www.infoq.com/articles/no-reliable-messaging>

第8章 并行程序调试

并行程序调试要比串行程序复杂得多，但幸运的是，现代IDE开发环境可以在一定程度上缓解并发程序调试的难度。在本章中，我想简单介绍一下有关并行程序调试的一些技巧和经验。

8.1 准备实验样本

为了方便讲解，我们定义一个简单的类，作为实验样本：

```
01 public class UnsafeArrayList {
02     static ArrayList al=new ArrayList();
03     static class AddTask implements Runnable{
04         @Override
05         public void run() {
06             try {
07                 Thread.sleep(100);
08             } catch (InterruptedException e) {}
09             for(int i=0;i<1000000;i++)
10                 al.add(new Object());
11         }
12     }
13     public static void main(String[] args) throws InterruptedE
14         Thread t1=new Thread(new AddTask(),"t1");
15         Thread t2=new Thread(new AddTask(),"t2");
16         t1.start();
17         t2.start();
18         Thread t3=new Thread(new Runnable(){
19             @Override
20             public void run() {
21                 while(true){
```

```
22         try {
23             Thread.sleep(1000);
24         } catch (InterruptedException e) {}
25     }
26 }
27 }, "t3");
28 t3.start();
29 }
30 }
```

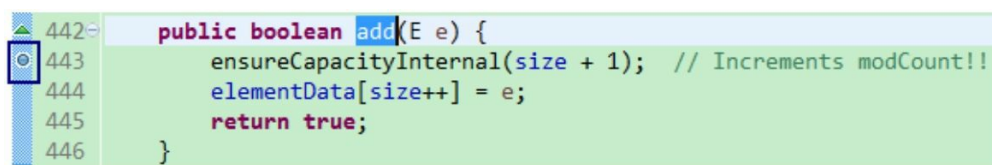
在这里，我使用的JDK版本为JDK8u5。

上述代码是在多线程下访问ArrayList，因此，是错误的写法。在这里，我们将使用调试，重现这个错误。

8.2 正式起航

在正式开始之前，先让我们熟悉一下Eclipse的调试环境。当你使用Eclipse调试Java程序时，当程序执行到断点处，默认情况下，当前的线程就会被挂起。

图8.1显示了在ArrayList.add()函数内部设置了一个断点：



```
442 public boolean add(E e) {  
443     ensureCapacityInternal(size + 1); // Increments modCount!!  
444     elementData[size++] = e;  
445     return true;  
446 }
```

图8-1 将断点设置在ArrayList.add()内

接着，以调试方式启动上面的代码，可以看到，程序会停留在系统第一次调用ArrayList.add()的地方，如图8.2所示。



```
UnsafeArrayList [Java Application]  
  geym.conc.ch8.UnsafeArrayList at localhost:23564  
    Thread [main] (Suspended (breakpoint at line 443 in ArrayList))  
      owns: URLClassPath (id=23)  
      owns: Object (id=24)  
      owns: Object (id=25)  
      ArrayList<E>.add(E) line: 443  
      URLClassPath.getLoader(int) line: 344  
      URLClassPath.getResource(String, boolean) line: 198  
      URLClassLoader$1.run() line: 364  
      URLClassLoader$1.run() line: 361  
      AccessController.doPrivileged(PrivilegedExceptionAction<T>, AccessControlContext) line: not available  
      Launcher$ExtClassLoader(URLClassLoader).findClass(String) line: 360  
      Launcher$ExtClassLoader(ClassLoader).loadClass(String, boolean) line: 424  
      Launcher$AppClassLoader(ClassLoader).loadClass(String, boolean) line: 411  
      Launcher$AppClassLoader.loadClass(String, boolean) line: 308  
      Launcher$AppClassLoader(ClassLoader).loadClass(String) line: 357  
      LauncherHelper.checkAndLoadMain(boolean, int, String) line: 495  
D:\tools\jdk8u5\bin\javaw.exe (2015年5月9日 下午1:02:19)
```

图8.2 断点阻止了程序的运行

在上图8.2中，可以看到主线程main停留在ArrayList.add()中，并且

显示了完整的调用堆栈。但很不幸的是，其实我们对主函数并没有太大兴趣，因为这些都是JDK内部的代码实现。目前，我们更关心的是在程序中t1和t2线程对ArrayList的调用。因此，我们会更希望忽略这些无关的调用。对于ArrayList这种非常常用的类来说，如果不加识别地进行断点设置，对系统的整个调试会变得异常痛苦。那么应该怎么办呢？

依托于Eclipse的强大功能，我们很容易实现这点。我们可以为这个断点设置一些额外属性，如图8.3所示。

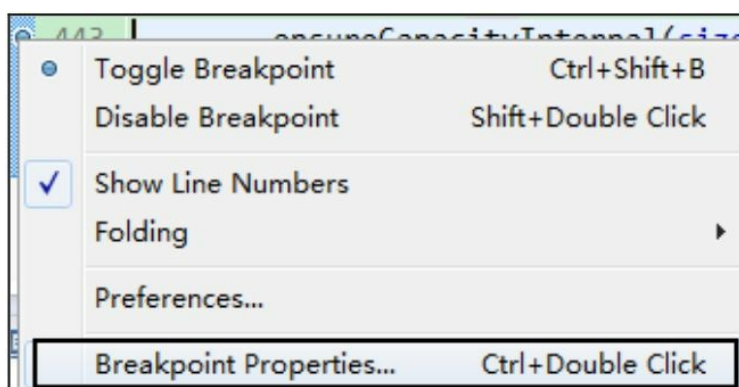


图8.3 设置断点属性

由于我们不希望主函数启动时被中断，因此在条件断点中指定断点条件是当前线程而不是主线程main，如图8.4所示，取得当前线程名称，并判断是否为主线程：

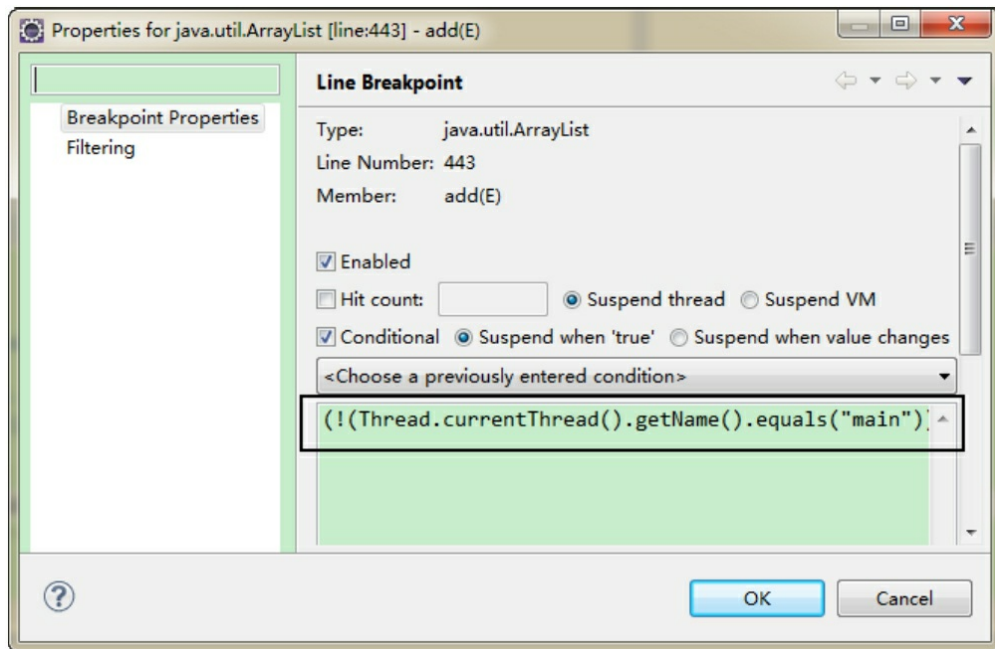


图8.4 设置条件断点

基于以上设置，再次执行调试这段代码，我们就可以调试t1和t2线程了，如图8.5所示。



图8.5 被中断的t1和t2

从这个调试窗口中可以看到，当前正在执行的几个线程，这里显示了t1、t2和t3。由于t3线程并没有使用ArrayList，因此，它处于Running状态，并保持一直执行。而t1和t2两个线程都在ArrayList.add()方法中被挂起。

如上图8.5所示，当前选中的是t2线程，如果我们进行单步操作，那么t2线程就会执行，而t1不会继续执行，除非，你手工选择t1并进行相应的操作。

8.3 挂起整个虚拟机

在这里，我还想提一个比较重要的功能。在默认情况下，当断点条件成立时，系统会挂起相关的线程，没有断点的线程会继续执行。在实际环境中，那些还在继续执行的线程可能会对整个调试产生不利的影响。为此，我们可以设置断点类型为挂起整个Java虚拟机，而不仅仅是挂起相关线程。如图8.6所示，改变这个断点的类型：

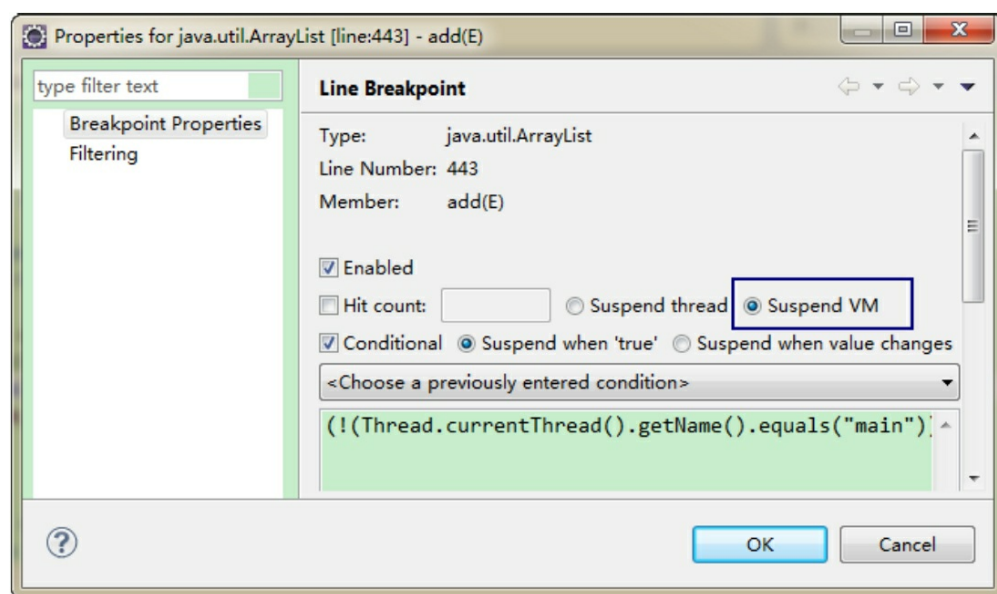


图8.6 设置断点类型为挂起整个虚拟机

当然，默认情况下，调试器只会挂起遇到断点的线程，如果你希望所有断点的模式都是挂起虚拟机而不是挂起线程，则还可以在Eclipse的全局配置中设置，如图8.7所示。

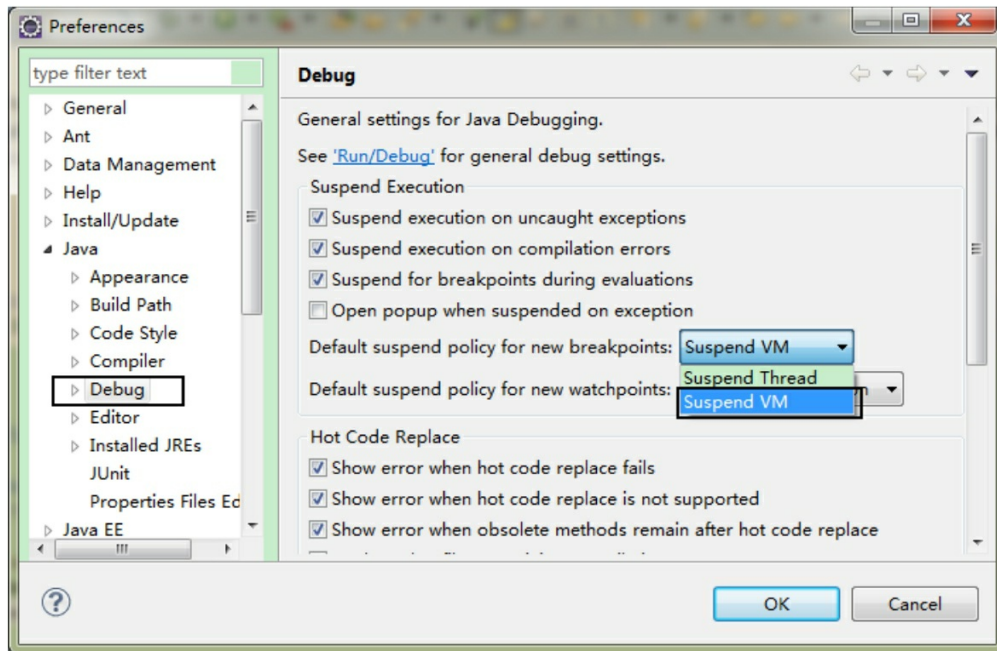


图8.7 设置断点模式行为为挂起虚拟机

在挂起虚拟机模式下，程序进入断点后的状态如图8.8所示。



图8.8 挂起虚拟机时的系统状态

可以看到，当前所有的线程全部处于挂起状态，不论当前线程是否接触到了断点。这种模式可以排除其他线程对被调试线程的干扰。当然，使用这种方法有时候会引起调试器或者虚拟机的一些问题，导致系统不能正常工作。

直接执行上述代码，很可能抛出类似下面的异常：

```
Exception in thread "t2" java.lang.ArrayIndexOutOfBoundsException  
    at java.util.ArrayList.add(ArrayList.java:444)  
    at geym.conc.ch8.UnsafeArrayList$AddTask.run(UnsafeArrayList.  
    at java.lang.Thread.run(Thread.java:745)
```

下面，就让我们用单步调试的方法，来重现这个异常吧！

8.4 调试进入ArrayList内部

首先，我们需要理解ArrayList的工作方式。在ArrayList初始化时，默认会分配10个数组空间。当数组空间消耗完毕后，ArrayList就会进行自动扩容。在每次add()操作时，系统总要事先检查一下内部空间是否满足所需的大小，如果不满足，就会扩容，否则就可以正常添加元素。

多线程共同访问ArrayList的问题在于：在ArrayList容量快用完时（只有1个可用空间），如果两个线程同时进入add()函数，并同时判断认为系统满足继续添加元素而不需要扩容，进而两者都不会进行扩容操作。之后，两个线程先后向系统写入自己的数据，那么必然有一个线程会将数据写到边界外，而产生这个ArrayIndexOutOfBoundsException。

基于上述原理，我们在ArrayList.add()函数中设置断点如图8.9所示。

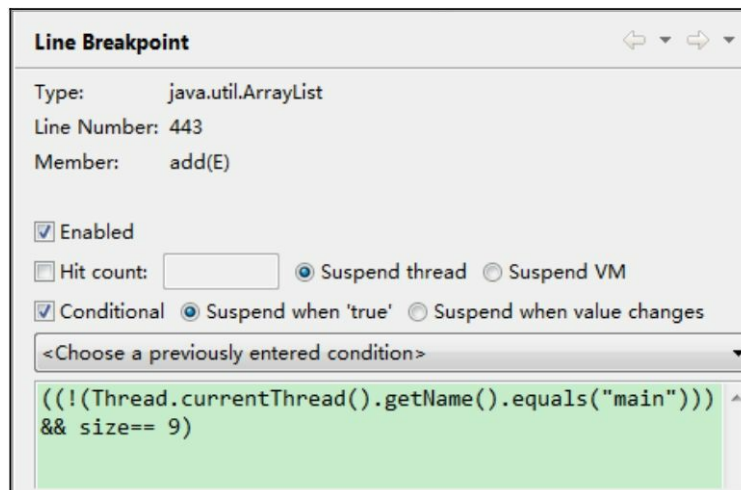


图8.9 ArrayList.add()的断点设置

这个断点意味着在非主线程中（这里就是t1和t2了），当进入add()

函数后，如果当前ArrayList的容量为9（当前的最大容量为10），则触发断点。之所以这么设置，是因为当容量没有饱和时，显然不会发生这个ArrayIndexOutOfBoundsException的问题，因此可以直接忽略这些情况。

接着，选中t1线程，让它进行容量检查，并让它停止在追加元素的语句前，如图8.10所示。

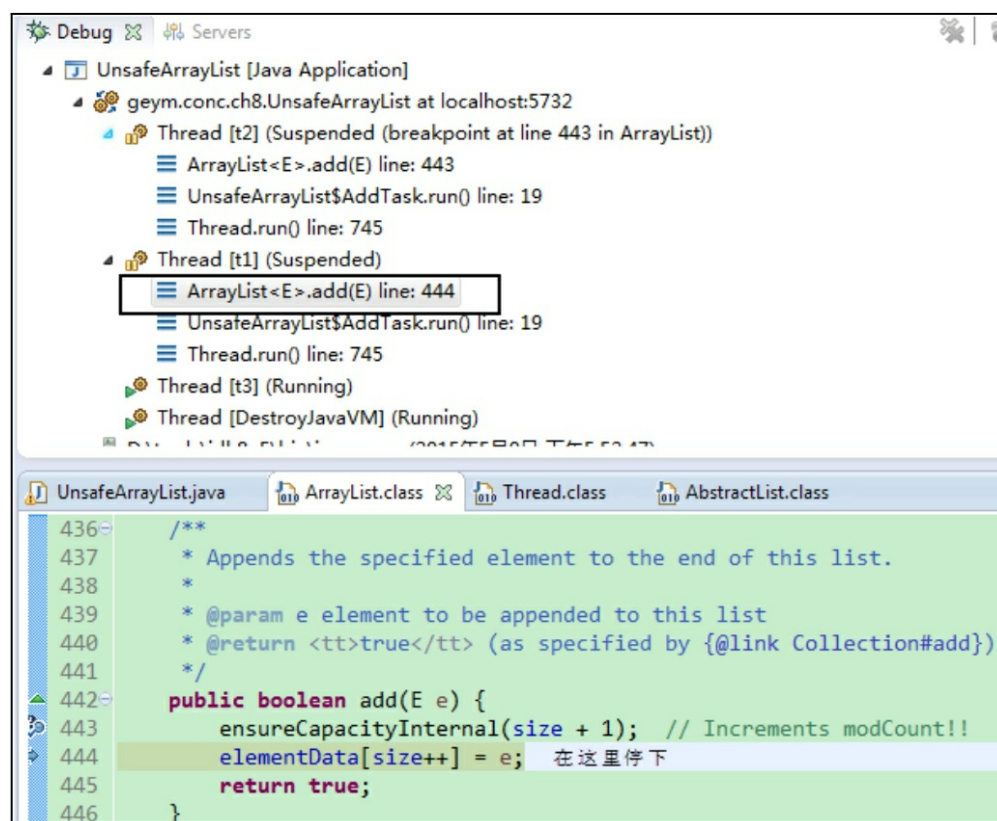


图8.10 t1线程完成容量检查

接着，在t1增加元素之前，选中t2线程，并让t2进入add()函数，完成进行容量检查，如图8.11所示。

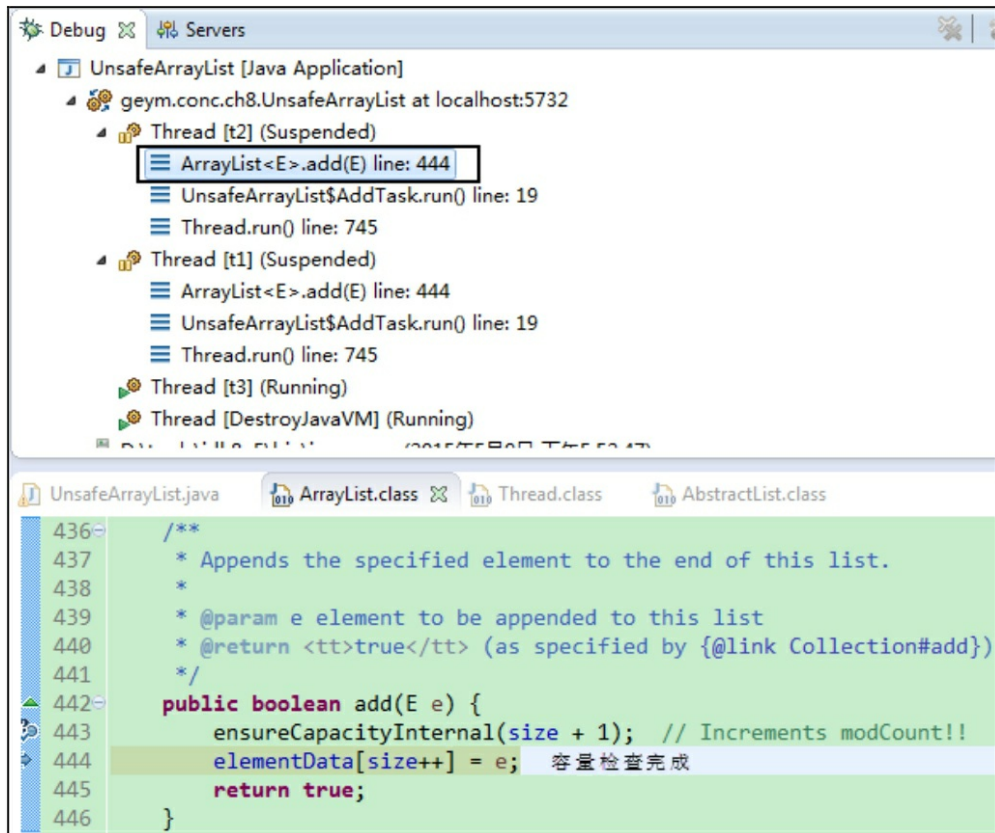


图8.11 t2完成容量检查

此时，t1和t2都认为ArrayList中的容量是满足它们的需求的，因此，它们都准备开始追加元素。让我们先选择t1，完成追加，如图8.12所示。

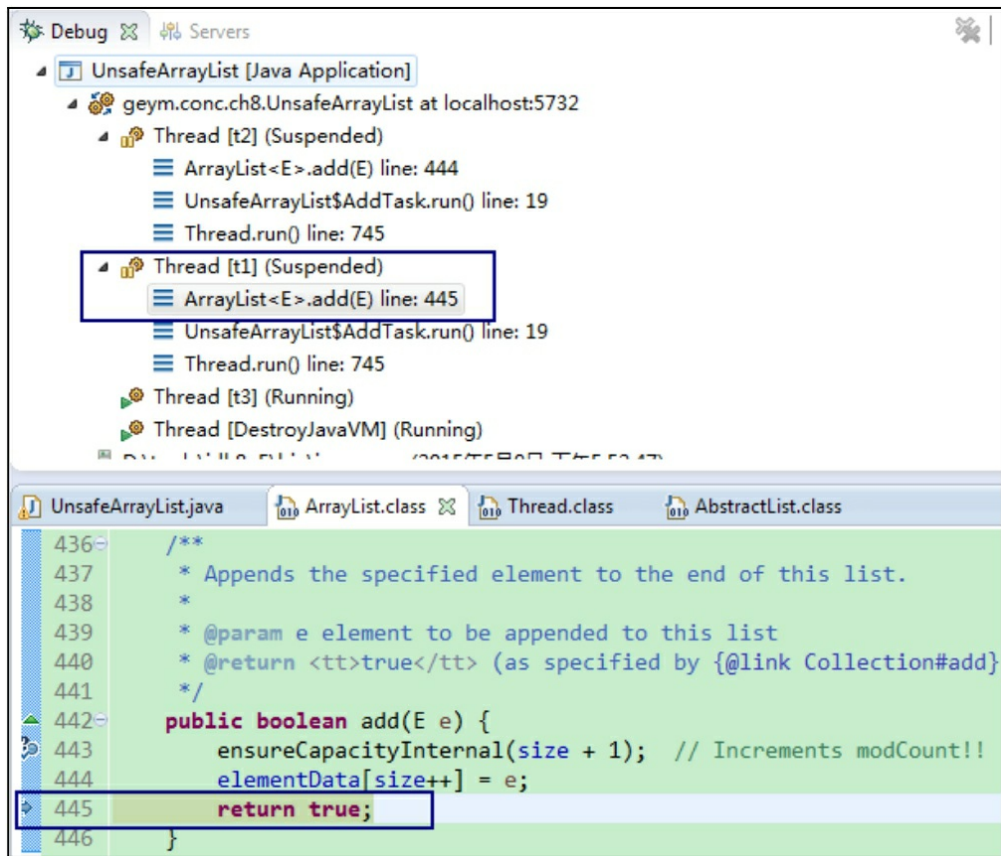


图8.12 t1完成元素追加

在t1追加完成后，t2并不知道数据空间实际上已经用完了。而之前的容量检查告诉t2，你可以继续追加元素，因此，t2还会义无反顾地继续执行后续追加操作。选择t2，让t2进行元素追加，此时，当t2试图向ArrayList追加元素时，追加操作并没有如我们预期一样完成，因为，此时，size的值已经超过了elementData的边界。如图8.13所示，可以看到ArrayIndexOutOfBoundsException异常位于t2线程中。

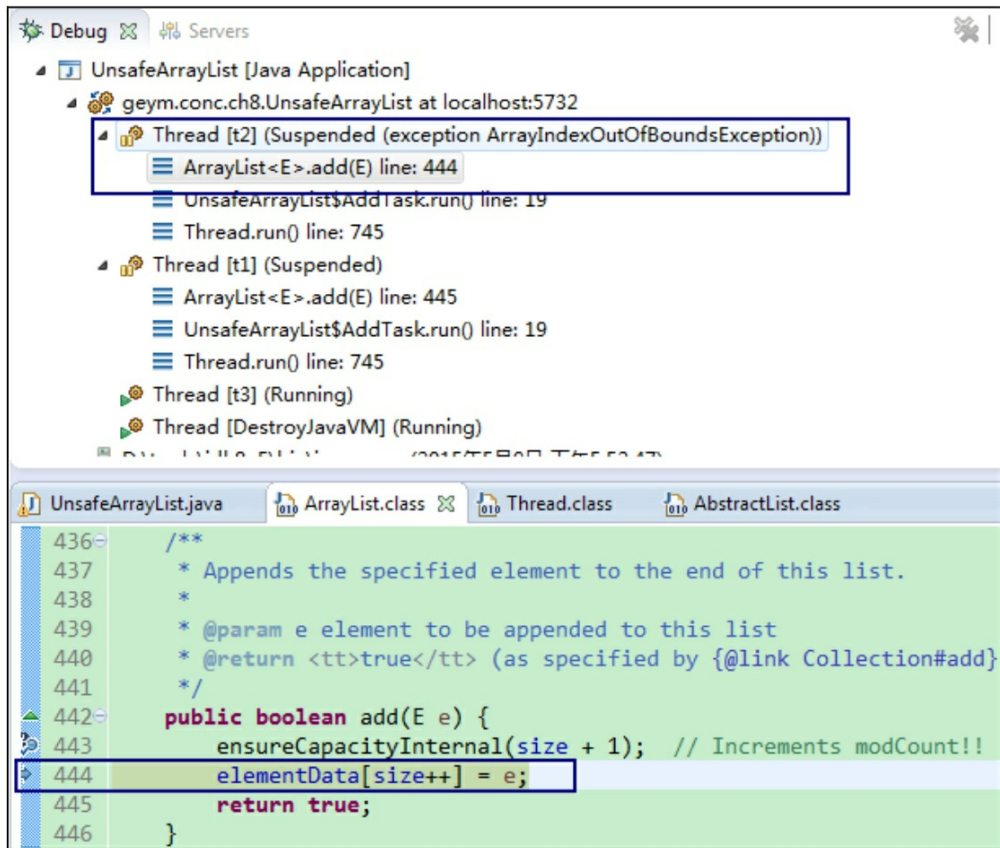


图8.13 t2线程发生异常

让t2继续往下执行的结果就是前文中那段异常信息，之后，t2线程就从线程列表中消失了（执行结束）。